When Vectors Break Down: Graph-Based RAG for Dense Enterprise Knowledge - Sam Julien, Writer

AI Engineer
243K subscribers
Subscribe

433 | Share | Ask | Download | ...

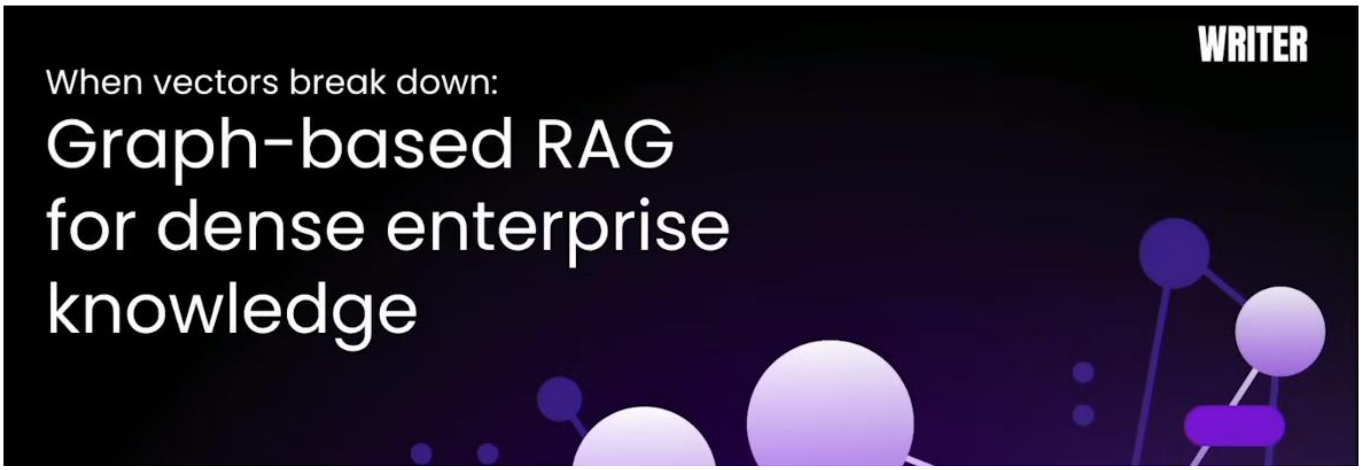
20,935 views Premiered Jul 22, 2025 AIEWF 2025 Complete Playlist

Enterprise knowledge bases are filled with "dense mapping," thousands of documents where similar terms appear repeatedly, causing traditional vector retrieval to return the wrong version or irrelevant information. When our customers kept hitting this wall with their RAG systems, we knew we needed a fundamentally different approach.

In this talk, I'll share Writer's journey developing a graph-based RAG architecture that achieved 86.31% accuracy on the RobustQA benchmark while maintaining sub-second response times, significantly outperforming vector approaches.

I'll survey the key techniques behind this performance leap and why graph-based approaches excel with complex enterprise information structures like product documentation, financial documents, and technical specifications that challenge traditional RAG systems. You'll learn about using specialized LLMs to build semantic relationships, how compression techniques efficiently handle concentrated enterprise data patterns, and how infusing key data points in the memory layer of the LLM lowers hallucination.

The presentation will provide practical insights into identifying when graph-based approaches make sense for your organization's specific data challenges, helping you make informed architectural decisions for your next enterprise RAG system.



The market is catching up: vector search is not enough for RAG at scale.

AI IN THE ENTERPRISE – 9 min read

Retrieval–augmented generation (RAG): *What* it is and *why* it's a hot topic for enterprise AI

KEVIN WEI | November 15, 2023

WRITER

INTUIT Qualcomm accenture salesforce

Dropbox Uber Vanguard FRANKLIN TEMPLETON

COMMVAULT Kenvue 6sense HubSpot

Writer Knowledge Graph:
>86% accuracy with <3% hallucinations

- GUI for data connectors & APIs
- 1. Specialized LLM to build graph
- 2. Retrieval-aware compression
- 3. Fusion-in-decoder
- 4. Transparent thought process
- Palmyra LLMs

Evaluation of RAG approaches: accuracy and response time

Pipeline	RobustQA Avg. score	Avg. response time(s)
Azure Cognitive Search Retriever + GPT-4 + Ada	72.36	>1.0s
Canopy (Pinecone)	59.61	>1.0s
LangChain + Pinecone + OpenAI	61.42	<0.8s
LangChain + Pinecone + Cohere	69.02	<0.6s
Llamaindex + Weaviate Vector Store - Hybrid Search	75.89	<1.0s
RAG Google Cloud Vertex AI Search + Bison	51.08	>0.8s
RAG Amazon SageMaker	32.74	<2.0s
Writer Knowledge Graph	**86.31**	**<0.6s**

Writer Knowledge Graph receives the highest RobustQA score (>86), with the fastest average response time (<0.6s)

See https://writer.com/engineering/rag-benchmark/

There are many ways to get the benefits of knowledge graphs in RAG!

How you get there is often just as valuable as the end result.

- Our journey to graph-based RAG
- What made our team successful

Our journey to graph-based RAG

⚠️ Caveat

🖼️ This is a sketch,
🏗️ Not a blueprint

Writer research

Enterprise-optimized models

Focus on developing more scalable, reliable, and transparent models specifically engineered for enterprise requirements

Practical evaluations

Development of model evaluation methodology that reflects real-world scenarios and risks

Domain-specific specialization

Research into applying AI systems in high-stakes industries

Retrieval & knowledge integration

Work on next-generation retrieval systems that safely and reliably connect language models with enterprise data

Real work → real insights → real users

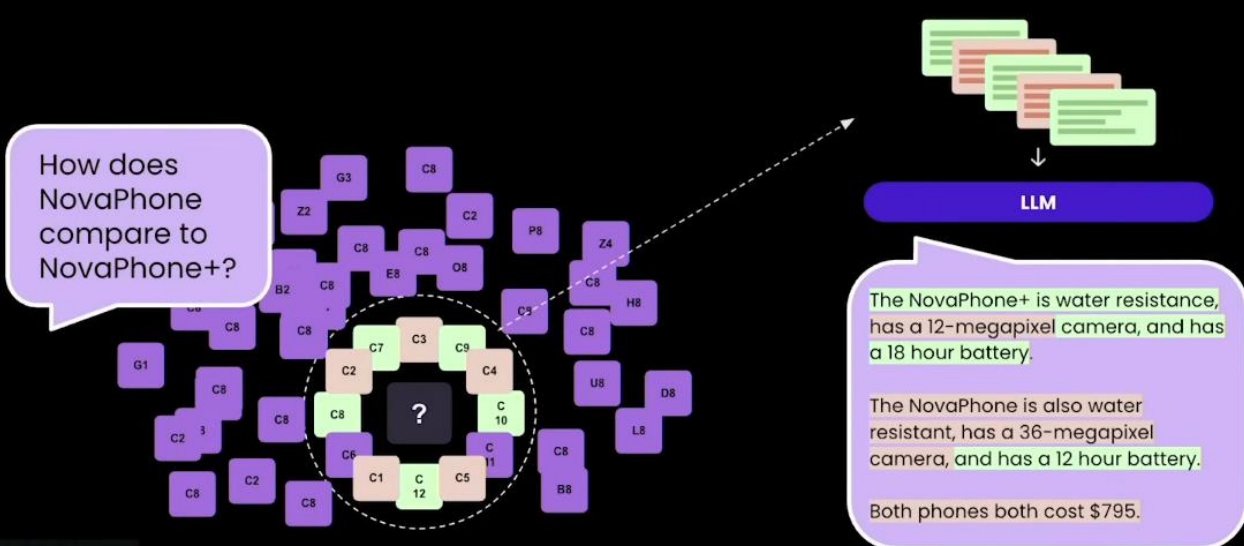
By embedding directly with product and customer teams, we translate customer and industry pain-points into model capabilities that actually make an impact. When AI research starts with real needs, it leads to:

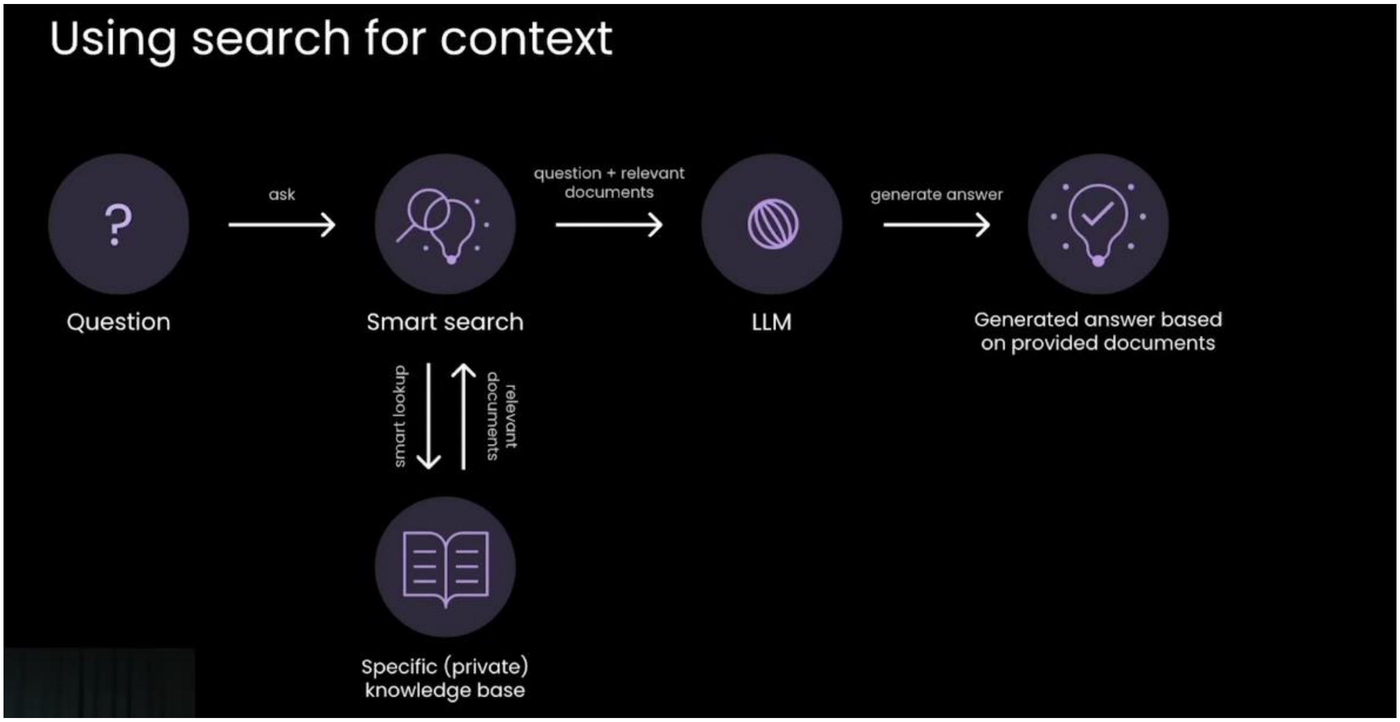
- Prioritizing capabilities that map to tangible outcomes
- Balanced focus between sophistication and practicality
- luation metrics to understand real-world performance
- ntification of potential risks and failures

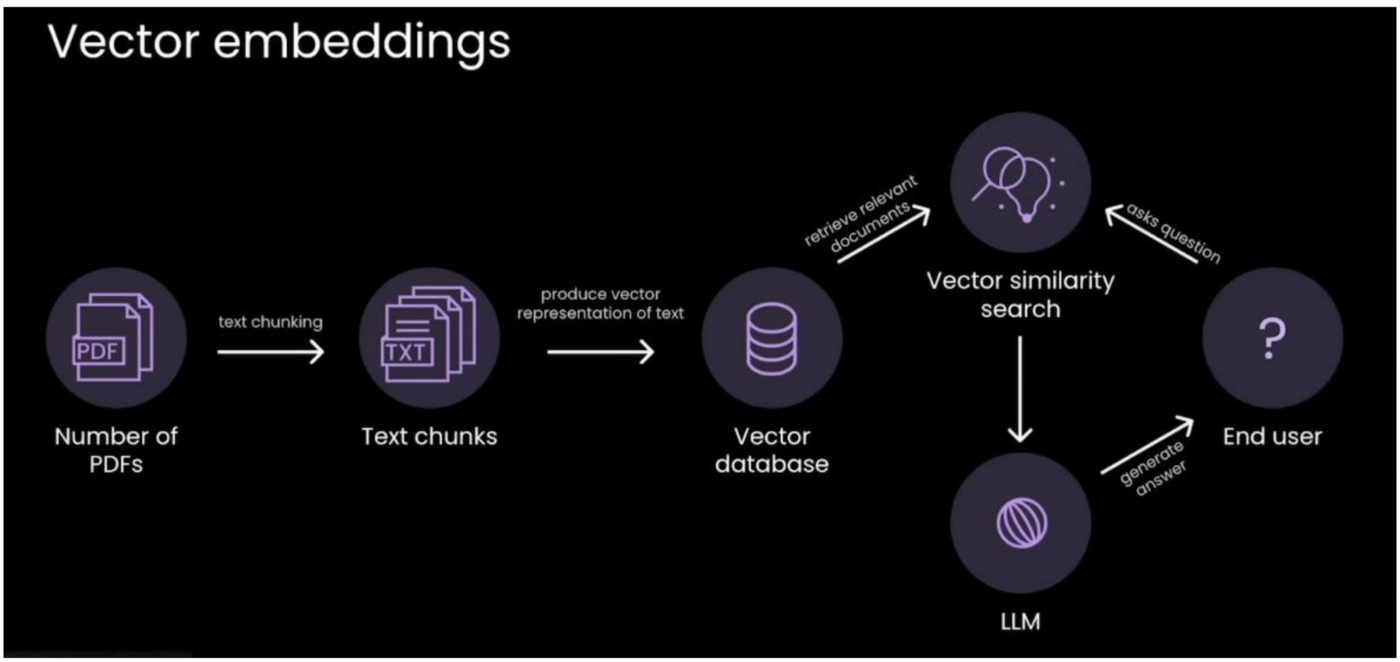
Focus on solving customer problems, not implementing specific solutions.

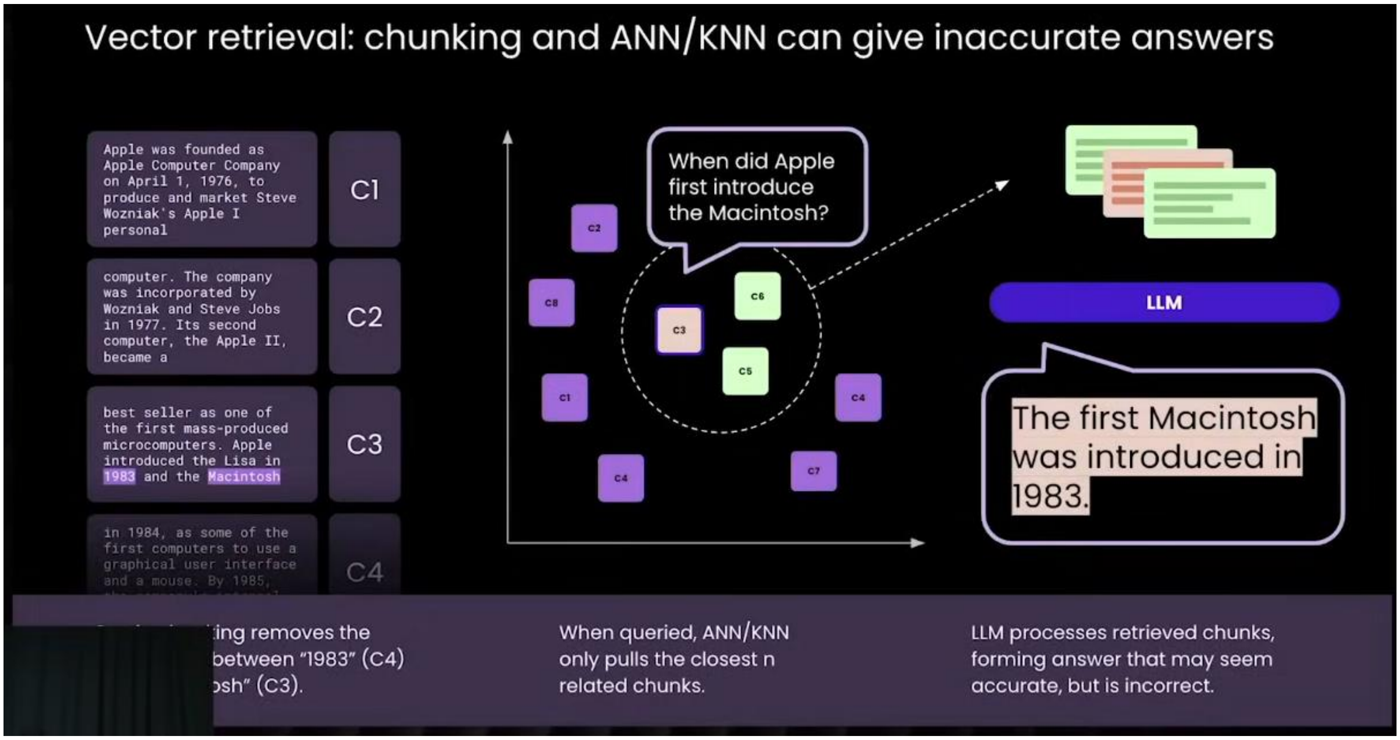
Don't chase what's hyped. Find the right solution for your customers.

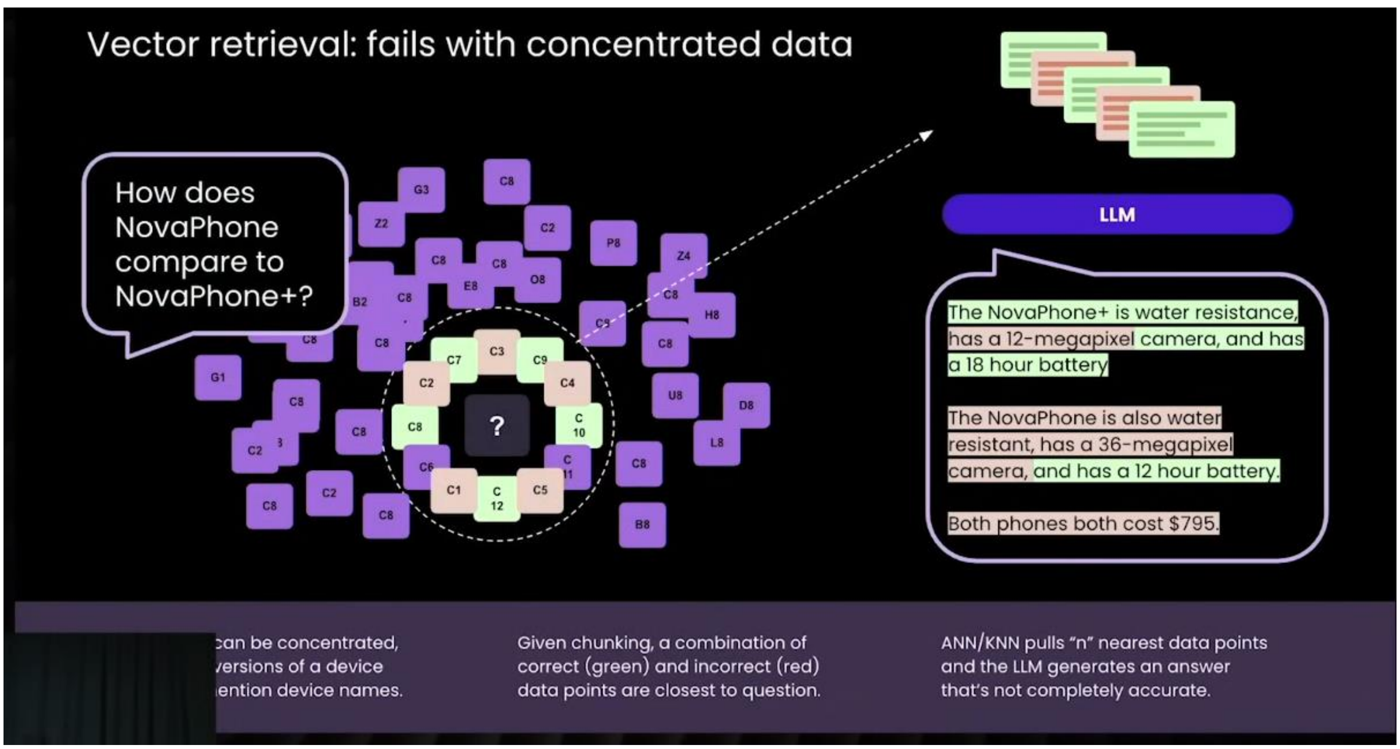
Enterprise data is dense, specialized, and massive

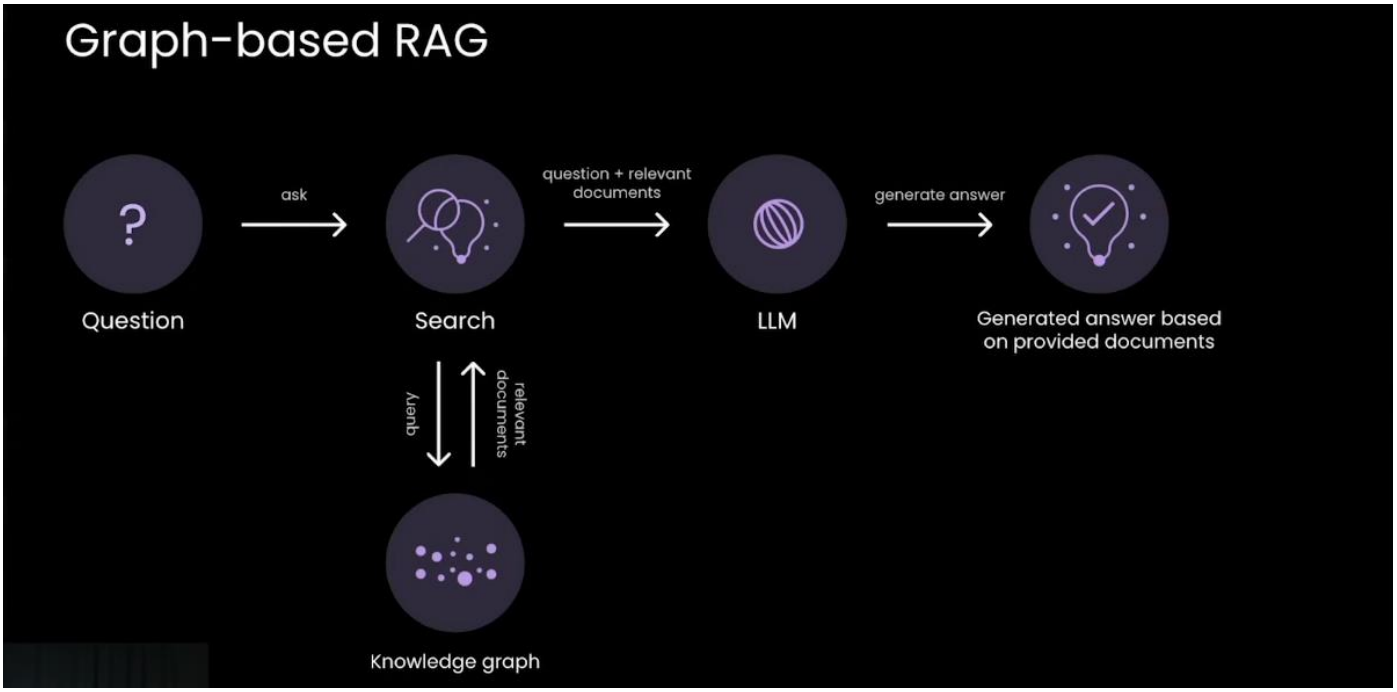


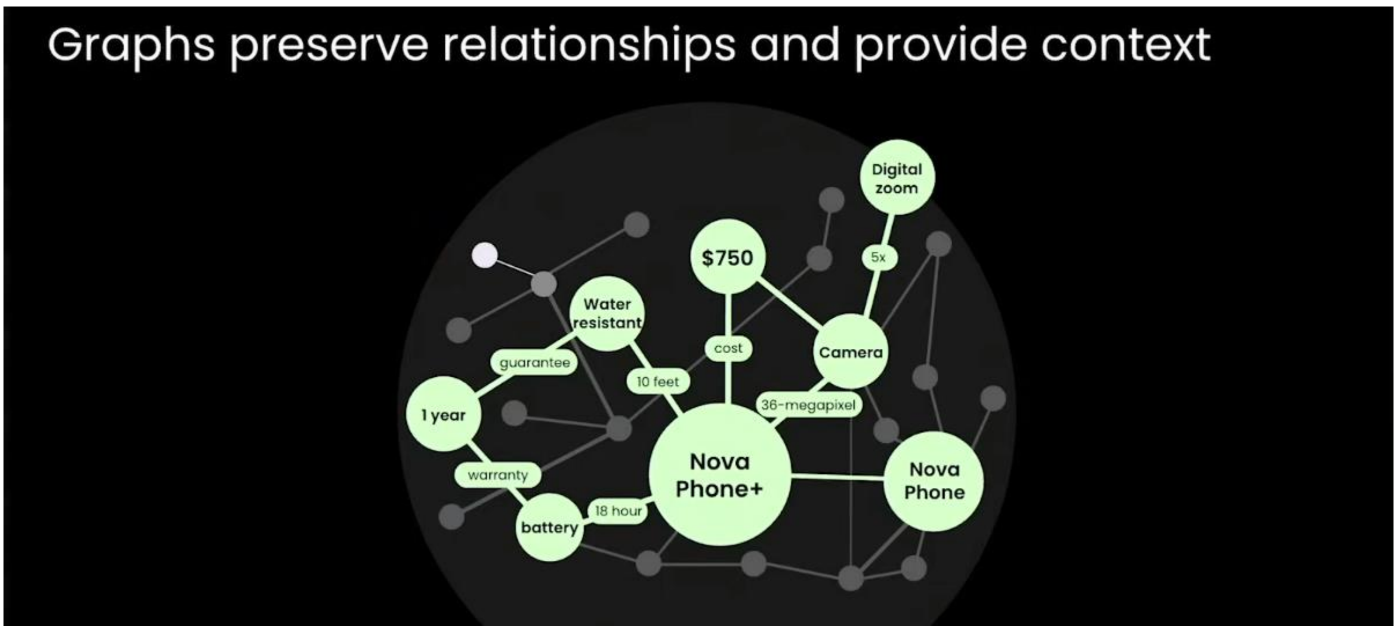

Using search for context
ask
question + relevant documents
generate answer
Question
Smart search
LLM
Generated answer based on provided documents
smart lookup
relevant documents
Specific (private) knowledge base


Vector embeddings
retrieve relevant documents
asks question
Vector similarity search
text chunking
produce vector representation of text
PDF
TXT
Number of PDFs
Text chunks
Vector database
End user
generate answer
LLM

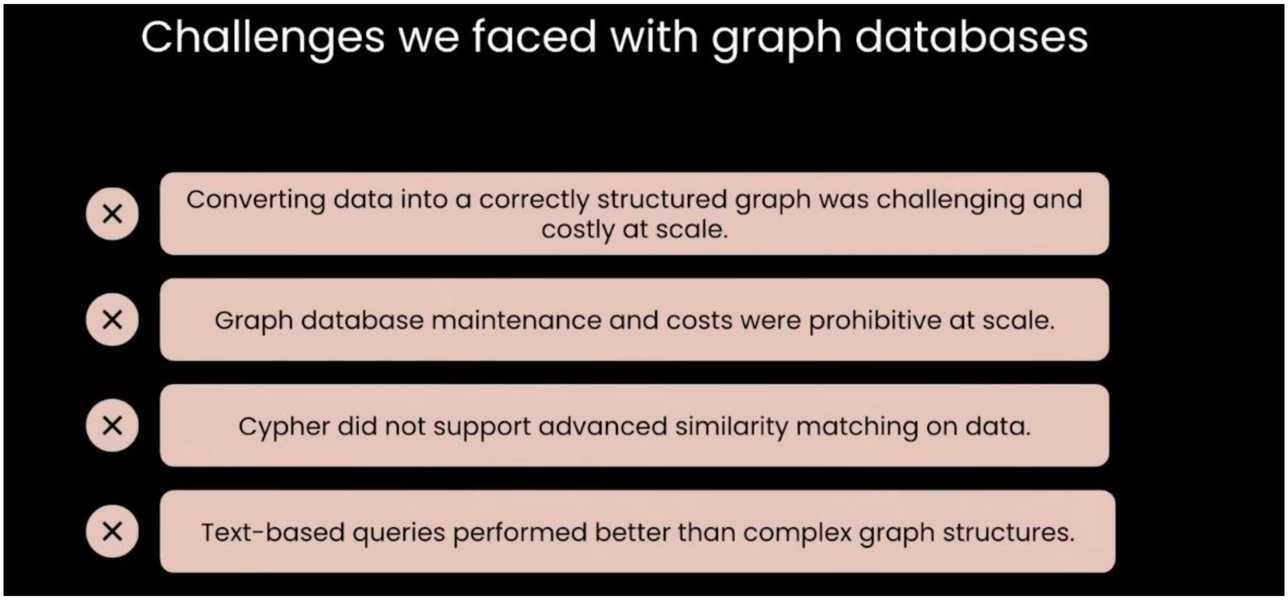

Vector retrieval: chunking and ANN/KNN can give inaccurate answers
Apple was founded as Apple Computer Company on April 1, 1976, to produce and market Steve Wozniak's Apple I personal
C1
computer. The company was incorporated by Wozniak and Steve Jobs in 1977. Its second computer, the Apple II, became a
C2
best seller as one of the first mass-produced microcomputers. Apple introduced the Lisa in 1983 and the Macintosh
C3
in 1984, as some of the first computers to use a graphical user interface and a mouse. By 1985,
C4
When did Apple first introduce the Macintosh?
C2
C8
C3
C6
C5
C1
C4
C4
C7
LLM
The first Macintosh was introduced in 1983.
ing removes the
between "1983" (C4)
sh" (C3).
When queried, ANN/KNN only pulls the closest n related chunks.
LLM processes retrieved chunks, forming answer that may seem accurate, but is incorrect.

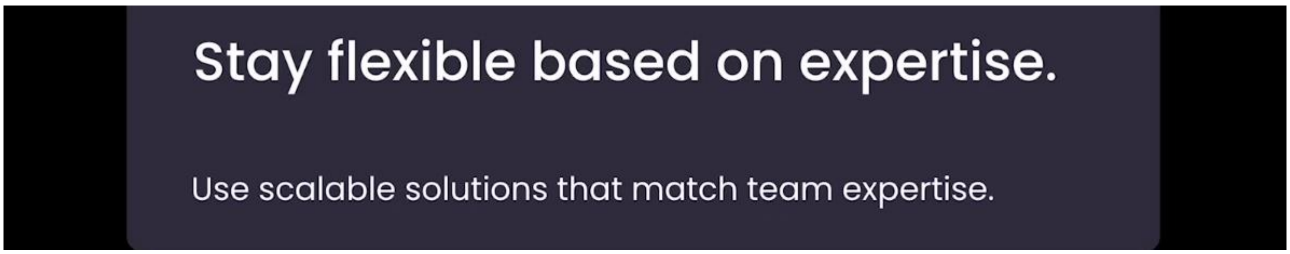

Vector retrieval: fails with concentrated data
How does NovaPhone compare to NovaPhone+?
?
C7
C3
C9
C2
C4
C8
C 10
C6
C1
C 12
C5
C 11
LLM
The NovaPhone+ is water resistance, has a 12-megapixel camera, and has a 18 hour battery
The NovaPhone is also water resistant, has a 36-megapixel camera, and has a 12 hour battery.
Both phones both cost $795.
can be concentrated,
versions of a device
ention device names.
Given chunking, a combination of correct (green) and incorrect (red) data points are closest to question.
ANN/KNN pulls "n" nearest data points and the LLM generates an answer that's not completely accurate.

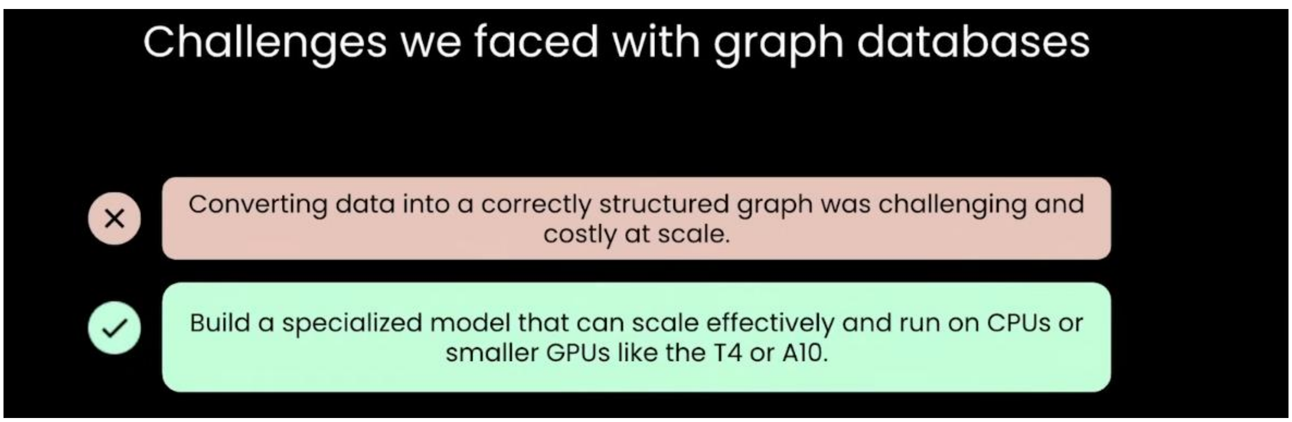




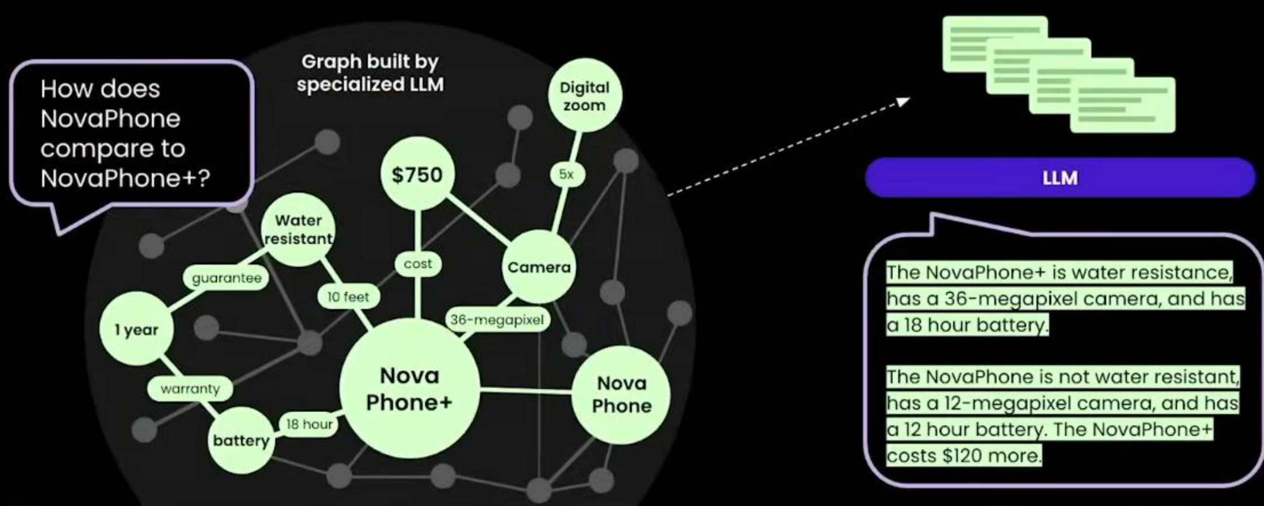
But, we ran into some challenges with graph databases as our customer needs scaled.


Challenges we faced with graph databases
Converting data into a correctly structured graph was challenging and costly at scale.
Graph database maintenance and costs were prohibitive at scale.
Cypher did not support advanced similarity matching on data.
Text-based queries performed better than complex graph structures.


Stay flexible based on expertise.
Use scalable solutions that match team expertise.


Challenges we faced with graph databases
Converting data into a correctly structured graph was challenging and costly at scale.
Build a specialized model that can scale effectively and run on CPUs or smaller GPUs like the T4 or A10.

Specialized LLM builds rich semantic relationships



[redacted] LLM trained to [redacted] graph structure [redacted] edges.

Rather than crude chunking, intelligently processes data based on context and maps semantic relationships.

Reliably retrieves correct data points for LLM to generate a completely accurate answer.

Challenges we faced with graph databases

✕ Graph database maintenance and costs were prohibitive at scale.

✓ Use scalable solutions that match team expertise.

Data points are stored in a Lucene-based search engine

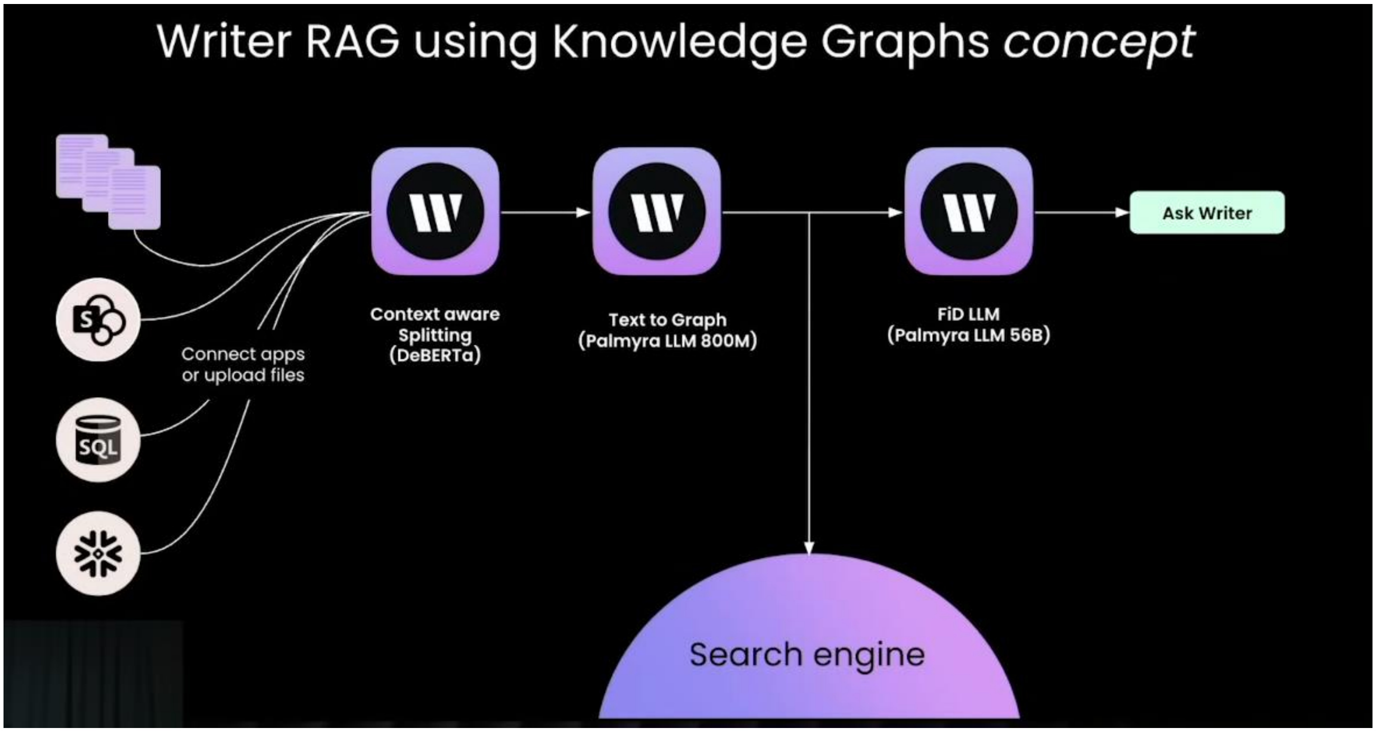
```
nodes = [
      "Nova Phone +",
      "Nova Phone",
      "Camera",
      "Digital zoom",
      "Battery",
      "Water resistance",
      "Warranty"
]
```

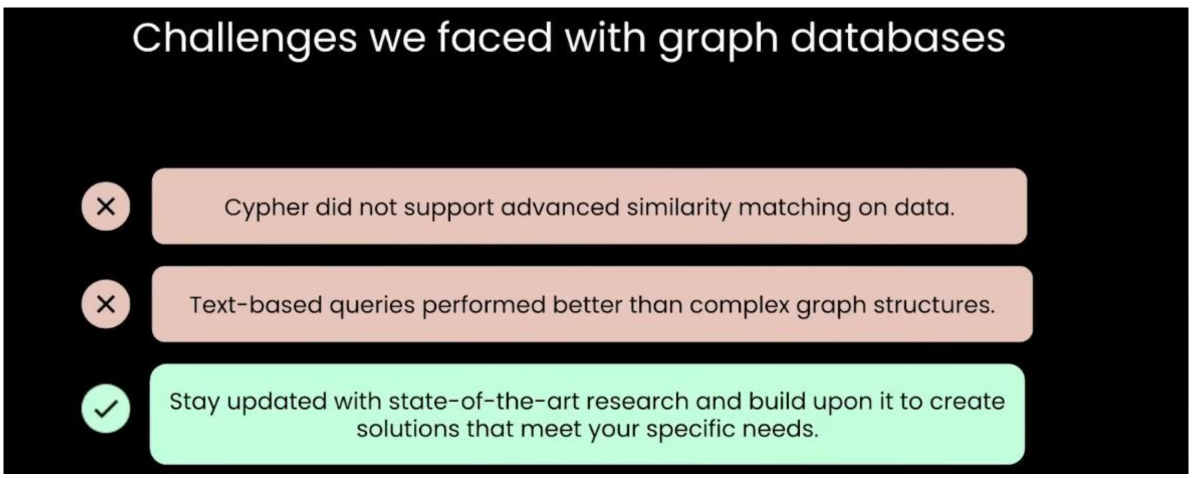
```
edges = [
      ("Nova Phone +", "Camera", {"relationship": "feature of"}),
      ("Nova Phone +", "$750", {"relationship": "cost of"}),
      ("Nova Phone +", "water resistant", {"relationship": "feature of"}),
      ("Nova Phone +", "battery", {"relationship": "feature of"}),
      ("Battery", "18 hour", {"relationship": "capacity of"}),
      ("Warranty", "1 year", {"relationship": "duration of"}),
      ("Water resistant", "10 feet", {"relationship": "limit of"}),
      ("Camera", "36-megapixel", {"relationship": "capability of"}),
      ("Digital zoom", "5x", {"relationship": "capability of"})
]
```

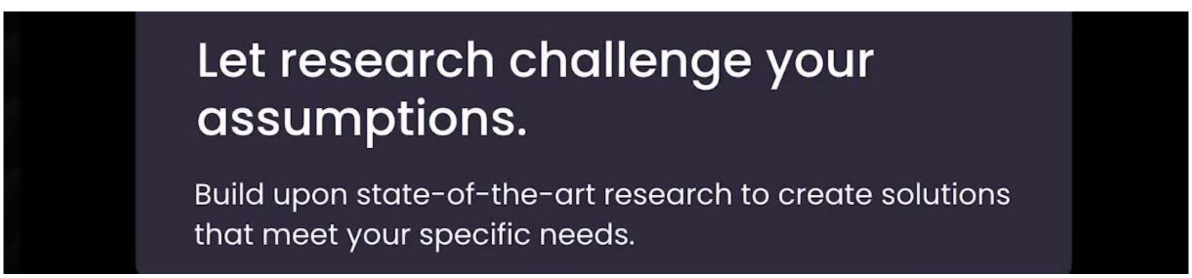
[redacted] re is converted

Uses search engine for storage rather than graph database.

Can easily handle large amounts of data without performance or speed degradation.


Writer RAG using Knowledge Graphs *concept*
Connect apps or upload files
Context aware Splitting (DeBERTa)
Text to Graph (Palmyra LLM 800M)
FiD LLM (Palmyra LLM 56B)
Ask Writer
Search engine


Challenges we faced with graph databases
Cypher did not support advanced similarity matching on data.
Text-based queries performed better than complex graph structures.
Stay updated with state-of-the-art research and build upon it to create solutions that meet your specific needs.


Let research challenge your assumptions.
Build upon state-of-the-art research to create solutions that meet your specific needs.

RAG didn't start as prompt + context + questions

Cornell University

arXiv > cs > arXiv:2005.11401

Computer Science > Computation and Language

[Submitted on 22 May 2020 (v1), last revised 12 Apr 2021 (this version, v4)]

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler,
Sebastian Riedel, Douwe Kiela

Large pre-trained language models have been shown to store factual knowledge in their parameters, and achieve state-of-the-ar
However, their ability to access and precisely manipulate knowledge is still limited, and hence on knowledge-intensive tasks, thei
Additionally, providing provenance for their decisions and updating their world knowledge remain open research problems. Pre-t
to explicit non-parametric memory can overcome this issue, but have so far been only investigated for extractive downstream tas
for retrieval-augmented generation (RAG) -- models which combine pre-trained parametric and non-parametric memory for lang
memory is a pre-trained seq2seq model and the non-parametric memory is a dense vector index of Wikipedia, access
lations, one which conditions on the same retrieved passages across the whole generated sequence, the other can use

Architecture from original RAG paper

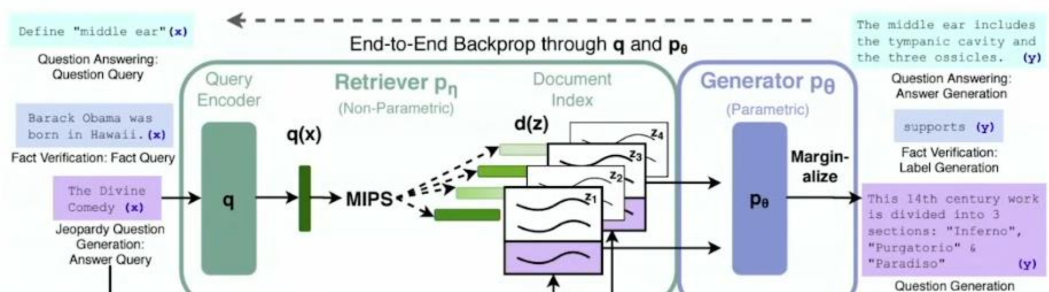


Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder* + *Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x, we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i. For final prediction y, we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

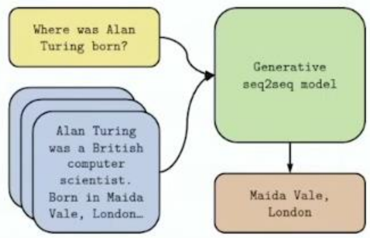
Fusion-in-decoder: Kind of an alternate RAG timeline

Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering

Gautier Izacard[1,2,3] Edouard Grave[1]
[1] Facebook AI Research, Paris
[2] ENS, PSL University, Paris
[3] Inria, Paris
gizacard|egrave@fb.com

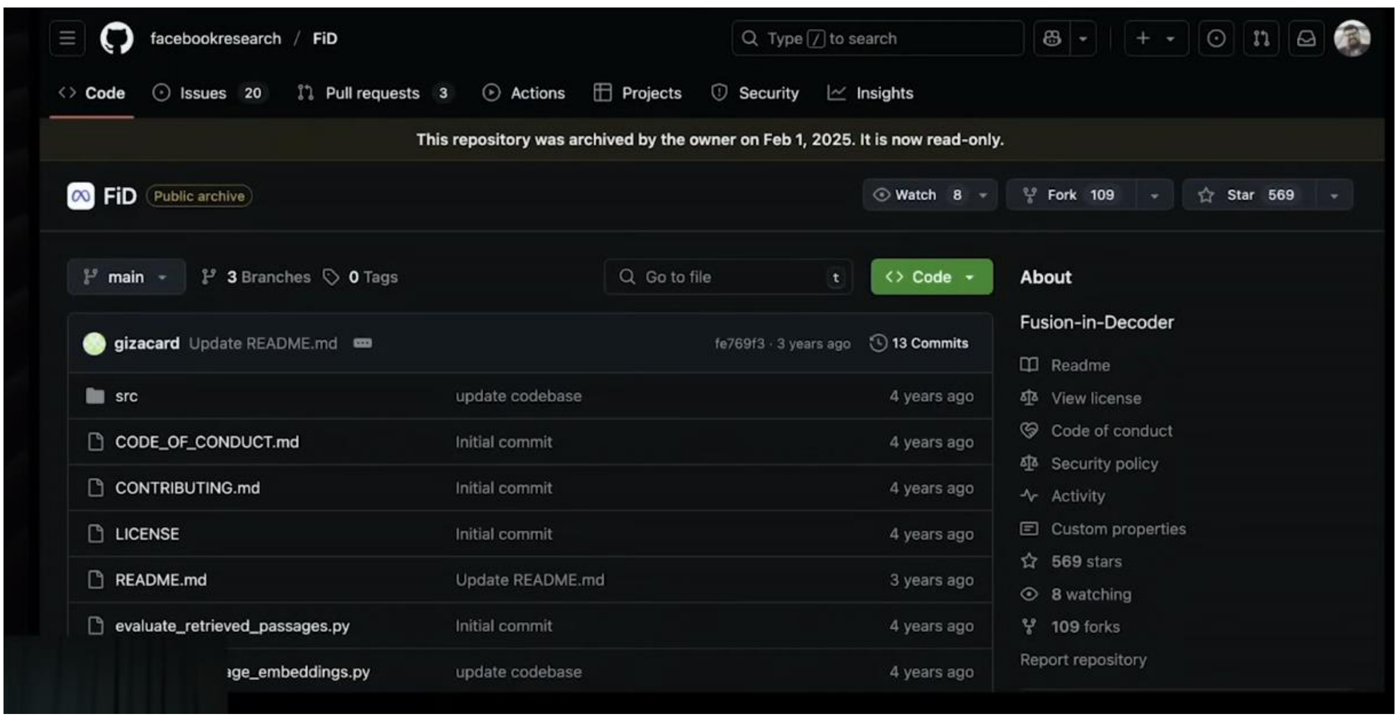
Abstract

Generative models for open domain question answering have proven to be competitive, without resorting to external knowledge. While promising, this approach requires to use models with billions of parameters, which are expensive to train and query. In this paper, we investigate how much these models can benefit from retrieving text passages, potentially containing evidence. We obtain state-of-the- … the Natural Questions and Triv-



- Partly motivated by improving upon the efficiency limitations of the original RAG approach while maintaining its retrieval-augmented benefits
- Process passages independently in the encoder (linear scaling) but jointly in the decoder (better evidence aggregation)
- Efficiency breakthroughs plus state-of-the-art performance!

rg/pdf/2007.01282

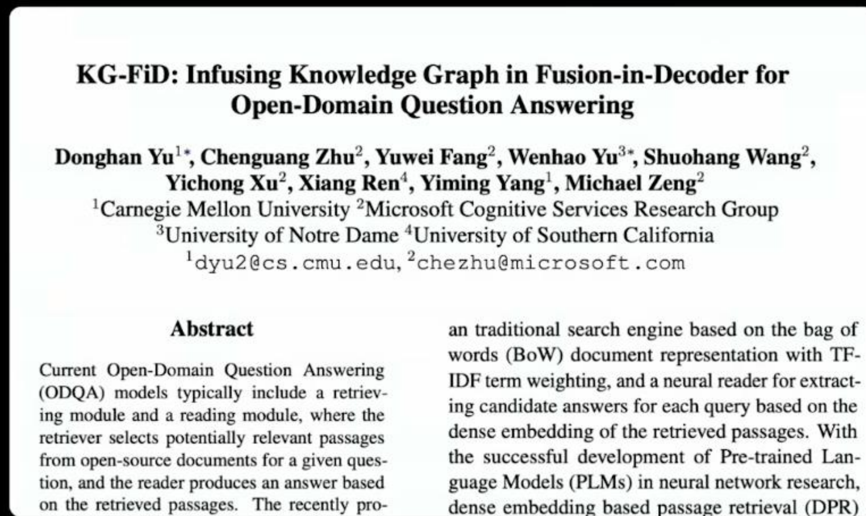


KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering

KG-FiD improves upon the original FiD model by using knowledge graphs to **understand relationships between retrieved passages**, rather than treating each passage independently.

The original Fusion-in-Decoder (FiD) was state-of-the-art but had two key issues:

1. **Independence assumption:** Passages were processed independently, ignoring relationships between them
2. **Efficiency bottleneck:** Processing ~100 passages per question was computationally ... e (6+ trillion operations ... tion)



KG-FiD: Infusing Knowledge Graph in Fusion-in-Decoder for Open-Domain Question Answering

Donghan Yu[1*], Chenguang Zhu[2], Yuwei Fang[2], Wenhao Yu[3*], Shuohang Wang[2], Yichong Xu[2], Xiang Ren[4], Yiming Yang[1], Michael Zeng[2]
[1]Carnegie Mellon University [2]Microsoft Cognitive Services Research Group
[3]University of Notre Dame [4]University of Southern California
[1]dyu2@cs.cmu.edu, [2]chezhu@microsoft.com

Abstract

Current Open-Domain Question Answering (ODQA) models typically include a retrieving module and a reading module, where the retriever selects potentially relevant passages from open-source documents for a given question, and the reader produces an answer based on the retrieved passages. The recently pro-

an traditional search engine based on the bag of words (BoW) document representation with TF-IDF term weighting, and a neural reader for extracting candidate answers for each query based on the dense embedding of the retrieved passages. With the successful development of Pre-trained Language Models (PLMs) in neural network research, dense embedding based passage retrieval (DPR)

https://arxiv.org/abs/2110.04330

KG-FiD Architecture

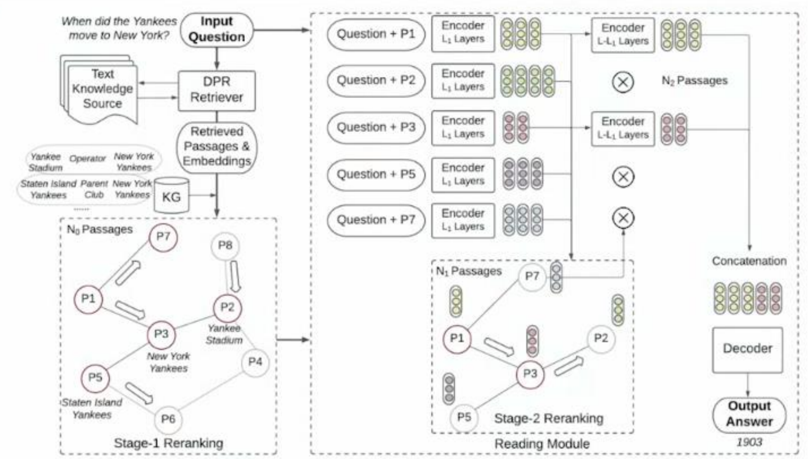
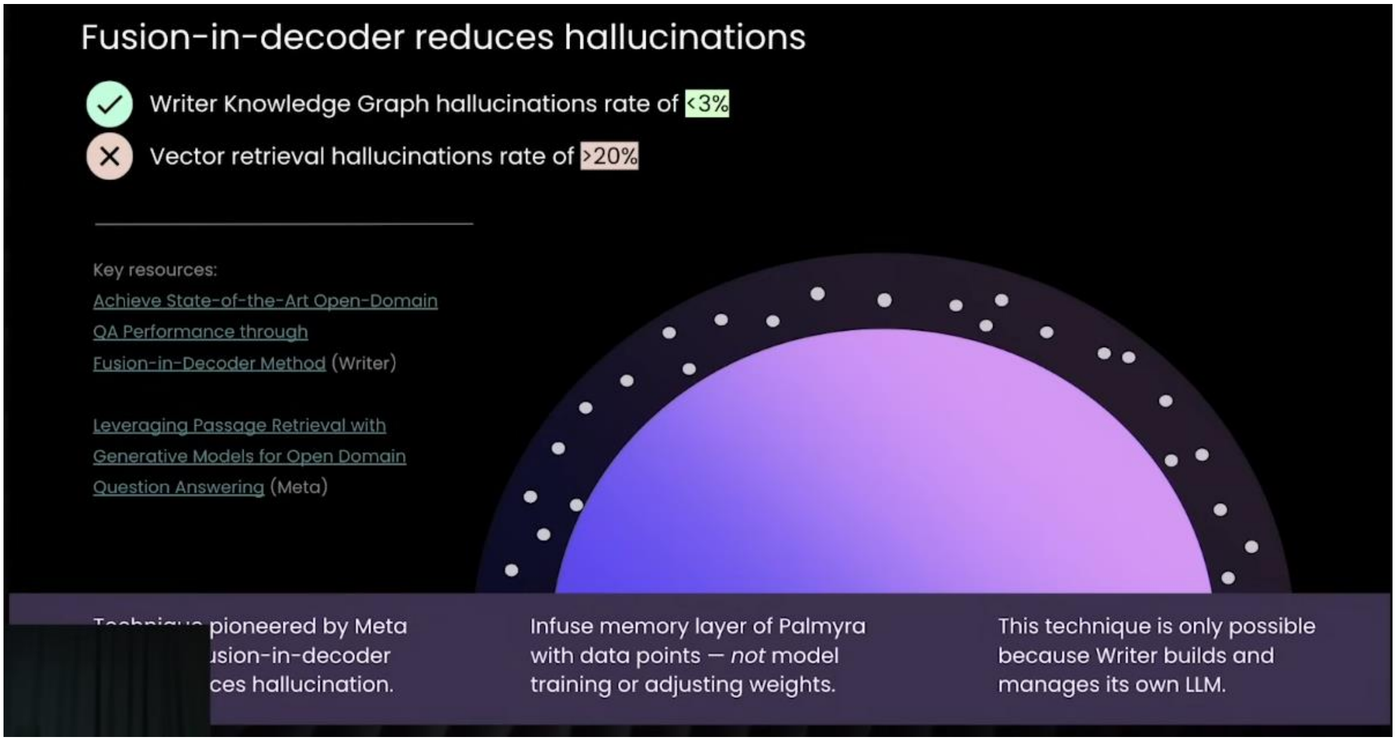
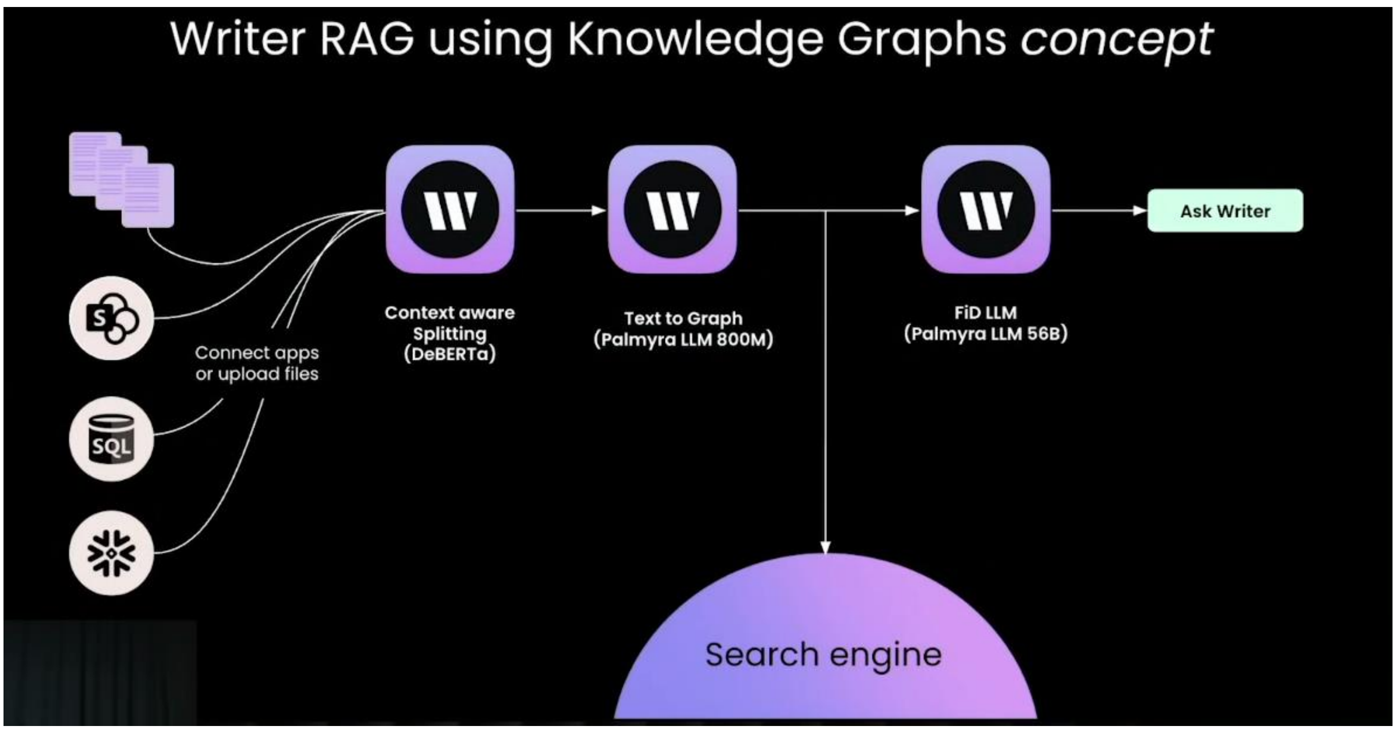


Figure 1: Overall Model Framework. Pi indicates the node of the passage originally ranked the i-th by the DPR retriever, with the article title below it. The left part shows passage retrieval by DPR, passage graph construction based on KG (Section 3.1) and stage-1 reranking (Section 3.2). The right part shows joint stage-2 reranking and answer generation in the reading module (Section 3.3 and 3.4).


Fusion-in-decoder reduces hallucinations
Writer Knowledge Graph hallucinations rate of <3%
Vector retrieval hallucinations rate of >20%
Key resources:
Achieve State-of-the-Art Open-Domain QA Performance through Fusion-in-Decoder Method (Writer)
Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering (Meta)
pioneered by Meta
usion-in-decoder
ces hallucination.
Infuse memory layer of Palmyra with data points — *not* model training or adjusting weights.
This technique is only possible because Writer builds and manages its own LLM.


Writer RAG using Knowledge Graphs *concept*
Connect apps or upload files
SQL
Context aware Splitting (DeBERTa)
Text to Graph (Palmyra LLM 800M)
FiD LLM (Palmyra LLM 56B)
Ask Writer
Search engine


But does it work?

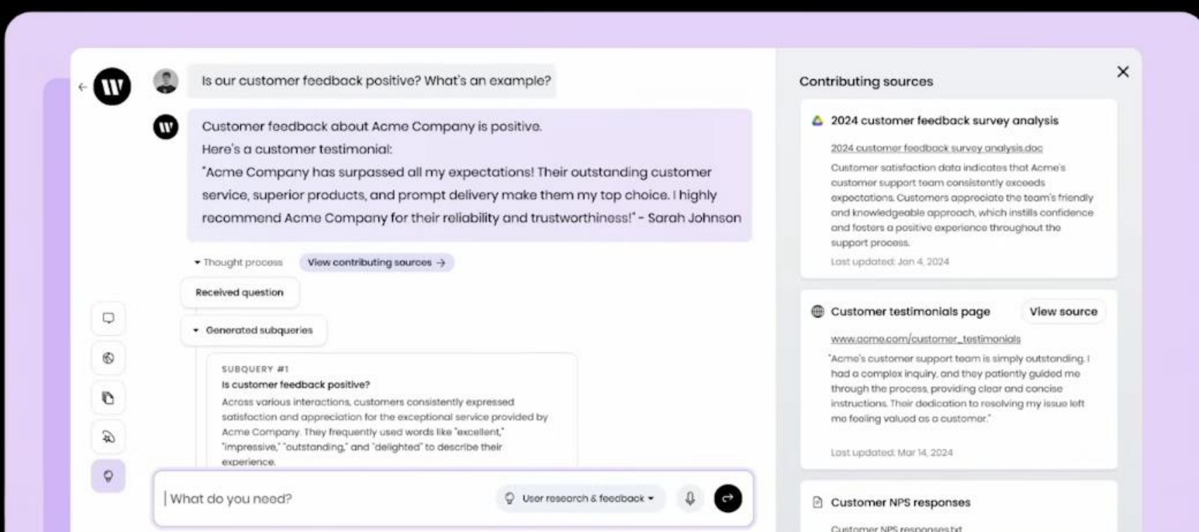
Benchmarking shows Writer Knowledge Graph achieves higher accuracy than vector retrieval

Retrieval systems	RobustQA Score	Response time
WRITER Knowledge Graph	86.31	<0.6s
LlamaIndex + Weaviate Vector Store + Weaviate Hybrid Search	75.89	<1.0s
Azure Cognitive Search Retriever + OpenAI + Ada	72.36	<1.0s
Langchain + Pinecone + Cohere	69.02	<0.6s
Langchain + Pinecone + OpenAI	61.42	<0.8s
Pinecone Canopy RAG + OpenAI	59.61	<1.0s
Google Vertex AI Search RAG + Bison	51.08	<0.8s
[redacted]eMaker RAG	32.74	<2.0s

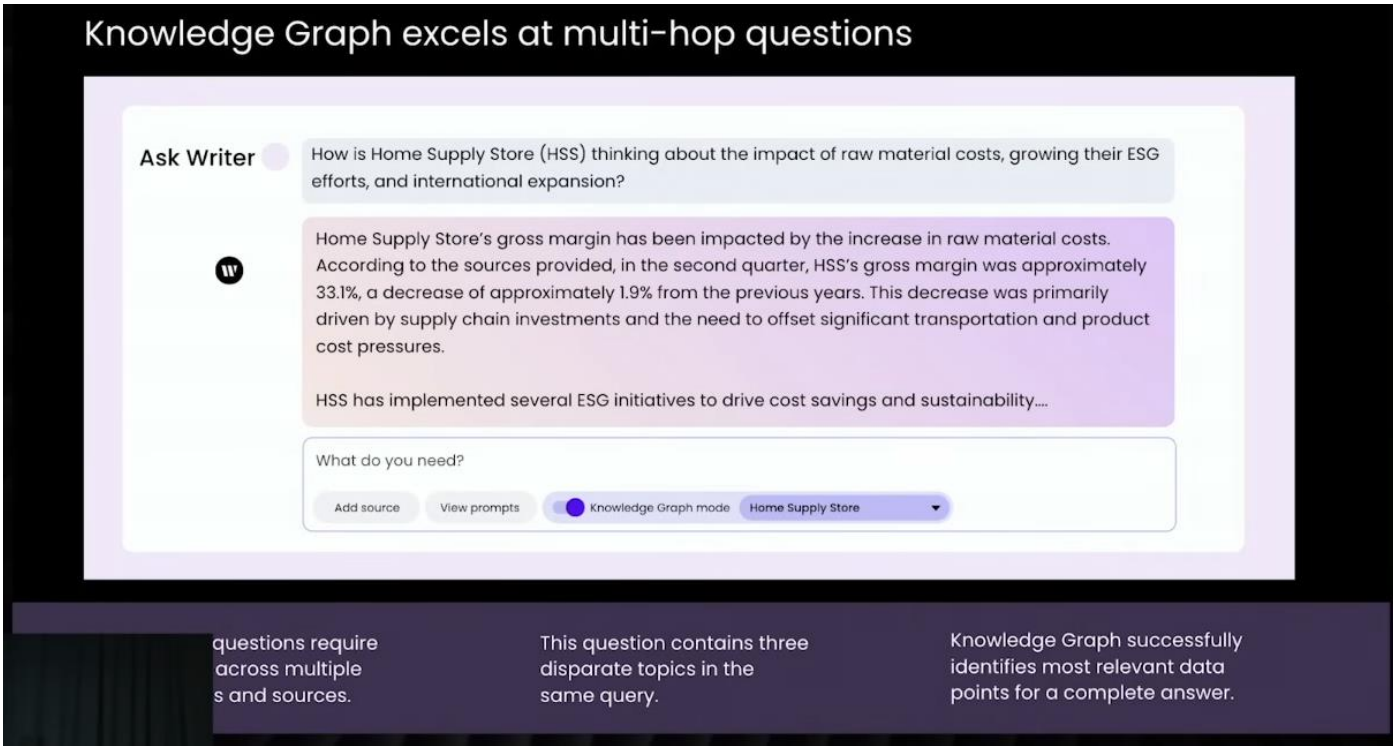
Writer Knowledge Graph achieves the **highest accuracy** and **fastest response time** compared to seven popular approaches to vector retrieval.

Source: Comparative Analysis of Retrieval Systems in the Real World

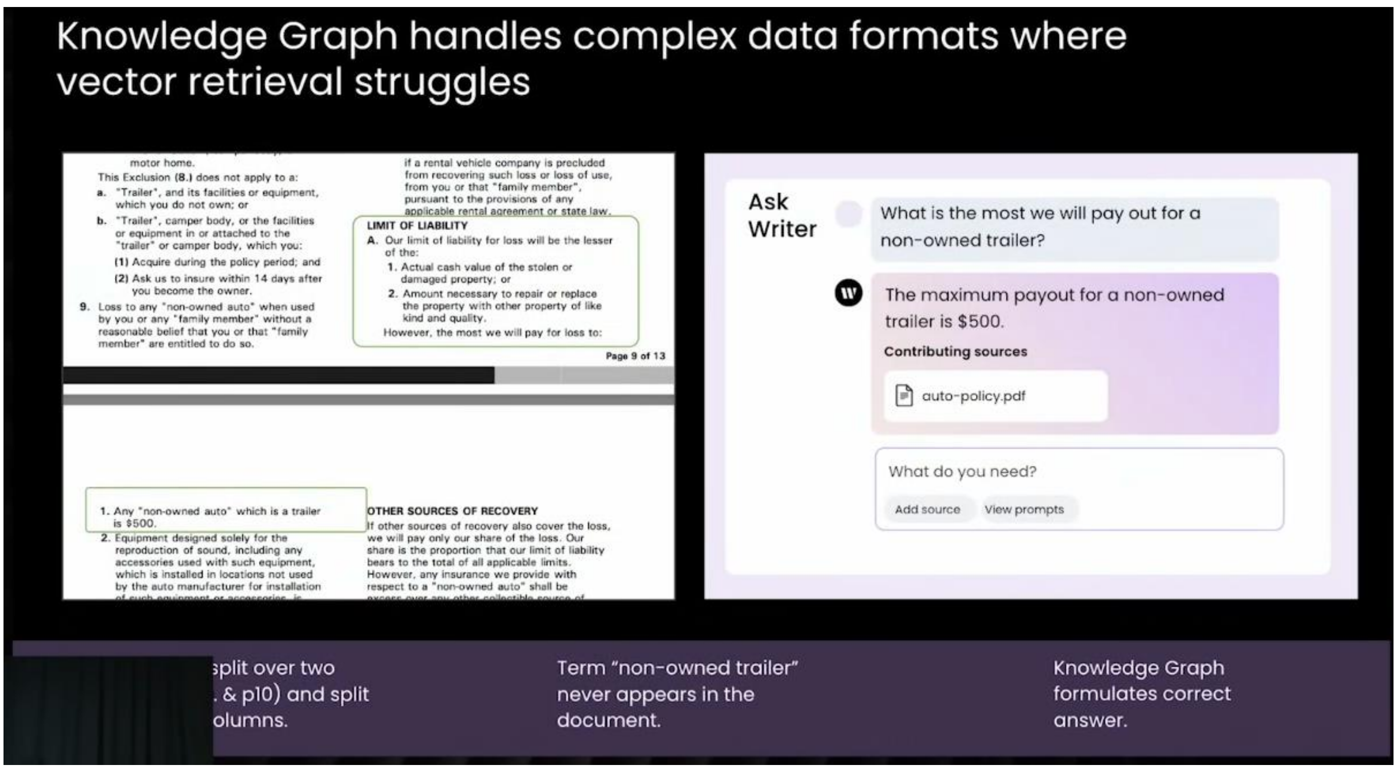
Knowledge Graph: transparent thought process for explainability



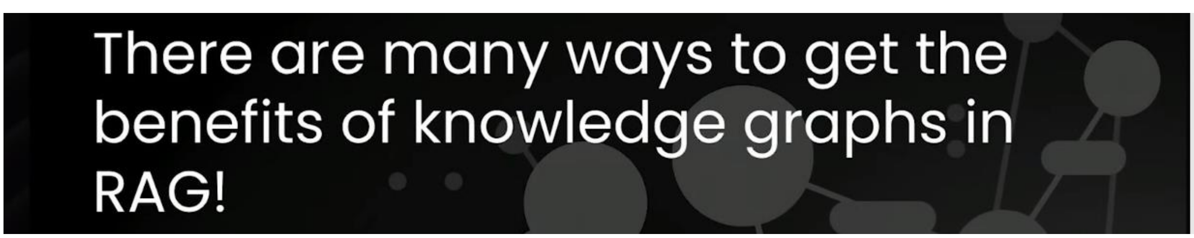
- Provides chain-of-thought in how answer was formulated
- Breaks complex queries into subquestions
- Shows sources and excerpts for answers



Knowledge Graph excels at multi-hop questions
Ask Writer
How is Home Supply Store (HSS) thinking about the impact of raw material costs, growing their ESG efforts, and international expansion?
Home Supply Store's gross margin has been impacted by the increase in raw material costs. According to the sources provided, in the second quarter, HSS's gross margin was approximately 33.1%, a decrease of approximately 1.9% from the previous years. This decrease was primarily driven by supply chain investments and the need to offset significant transportation and product cost pressures.
HSS has implemented several ESG initiatives to drive cost savings and sustainability....
What do you need?
Add source
View prompts
Knowledge Graph mode
Home Supply Store
questions require
across multiple
s and sources.
This question contains three disparate topics in the same query.
Knowledge Graph successfully identifies most relevant data points for a complete answer.



Knowledge Graph handles complex data formats where vector retrieval struggles
motor home.
This Exclusion (8.) does not apply to a:
a. "Trailer", and its facilities or equipment, which you do not own; or
b. "Trailer", camper body, or the facilities or equipment in or attached to the "trailer" or camper body, which you:
(1) Acquire during the policy period; and
(2) Ask us to insure within 14 days after you become the owner.
9. Loss to any "non-owned auto" when used by you or any "family member" without a reasonable belief that you or that "family member" are entitled to do so.
if a rental vehicle company is precluded from recovering such loss or loss of use, from you or that "family member", pursuant to the provisions of any applicable rental agreement or state law.
LIMIT OF LIABILITY
A. Our limit of liability for loss will be the lesser of the:
1. Actual cash value of the stolen or damaged property; or
2. Amount necessary to repair or replace the property with other property of like kind and quality.
However, the most we will pay for loss to:
Page 9 of 13
1. Any "non-owned auto" which is a trailer is $500.
2. Equipment designed solely for the reproduction of sound, including any accessories used with such equipment, which is installed in locations not used by the auto manufacturer for installation
OTHER SOURCES OF RECOVERY
If other sources of recovery also cover the loss, we will pay only our share of the loss. Our share is the proportion that our limit of liability bears to the total of all applicable limits. However, any insurance we provide with respect to a "non-owned auto" shall be
Ask Writer
What is the most we will pay out for a non-owned trailer?
The maximum payout for a non-owned trailer is $500.
Contributing sources
auto-policy.pdf
What do you need?
Add source
View prompts
split over two
. & p10) and split
olumns.
Term "non-owned trailer" never appears in the document.
Knowledge Graph formulates correct answer.



There are many ways to get the benefits of knowledge graphs in RAG!


How you get there is often just as valuable as the end result.


What made our team successful
1.
Focus on customer needs, not tools
Don't chase what's hyped. Find the right solution for your customers.
2.
Stay flexible based on expertise
Use scalable solutions that match team expertise.
3.
Let research challenge your assumptions
Build upon state-of-the-art research to create solutions that meet your specific needs.