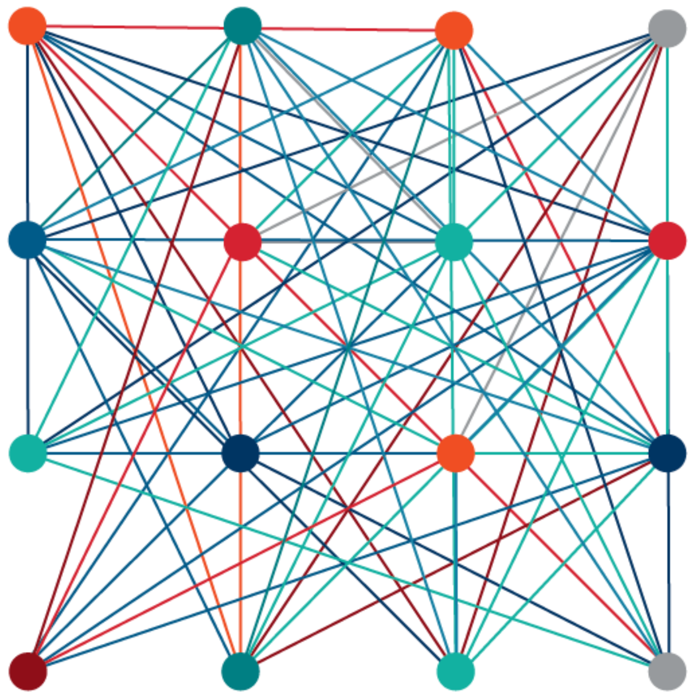
Graph RAG Varieties

KG+ AI Vision, architecture patterns, solution scoping, AI offerings and demonstrators

Knowledge graphs “connect the dots”!

ontotext



Talk to your data

to connect the dots

across data sources

and turn the complexity

of a global business

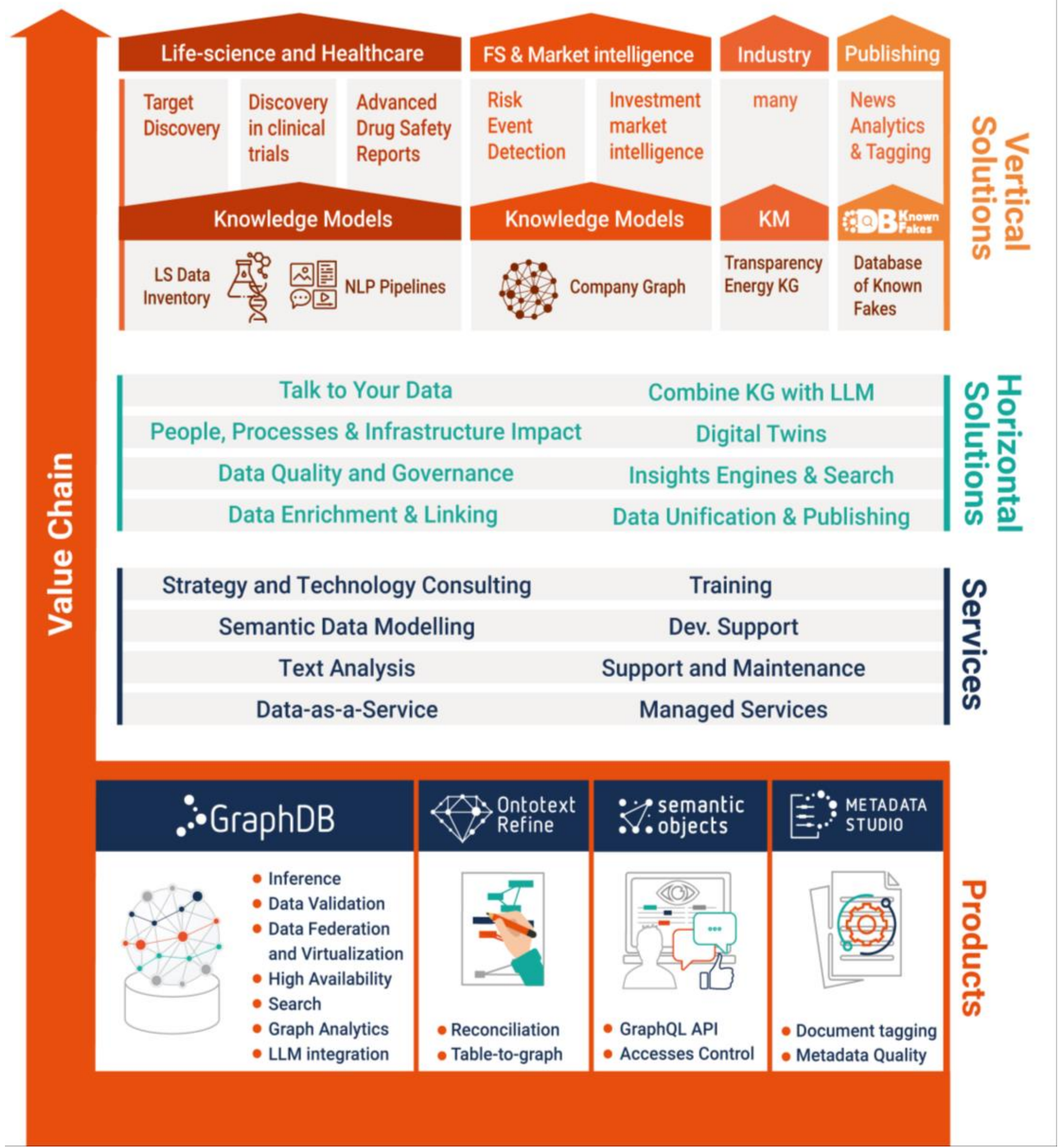
into a competitive advantage

Strong and diverse portfolio

ontotext

GraphDB™: The Most Robust and Versatile Graph Database

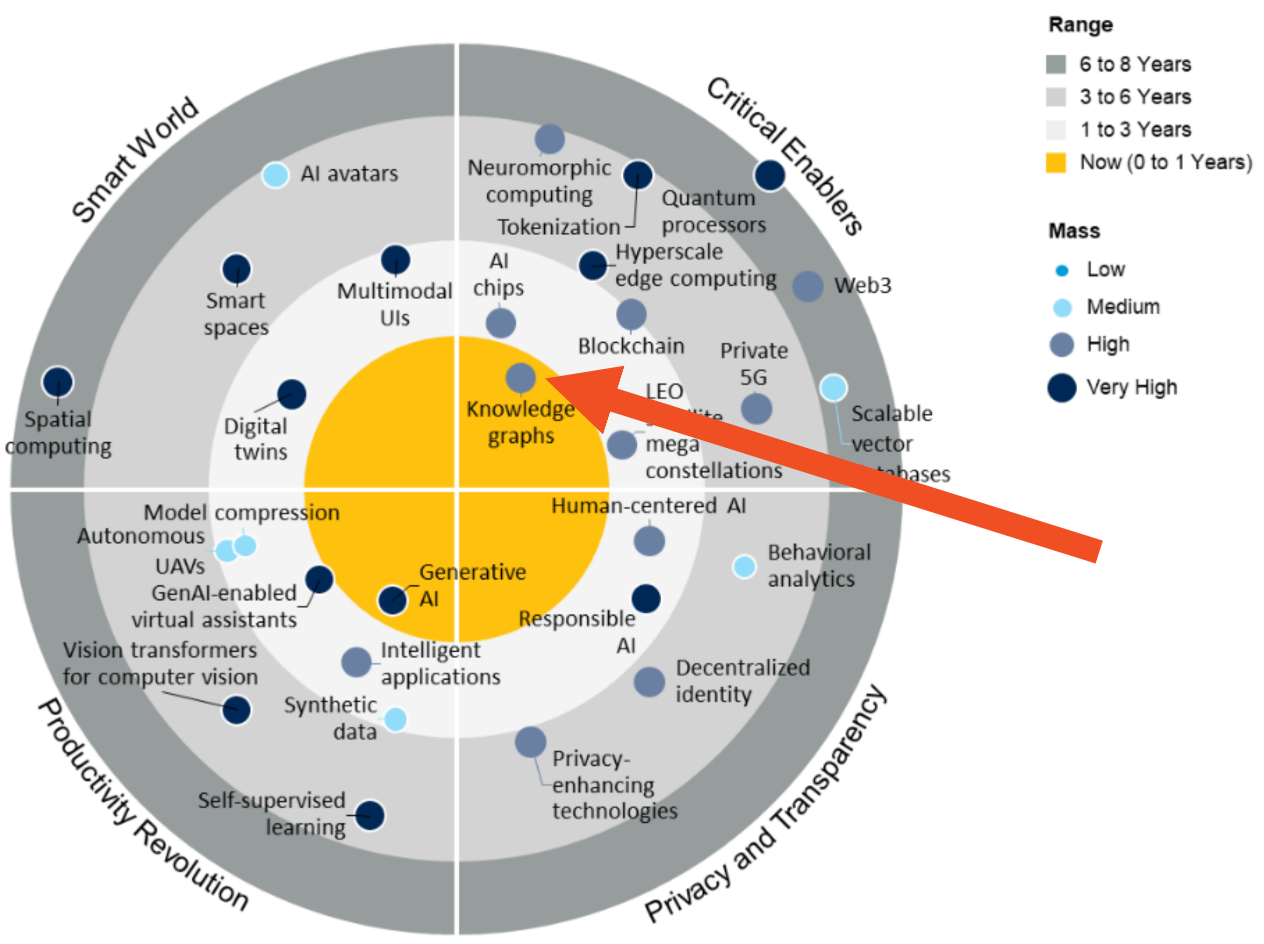
- Dependable performance across a variety of workloads
- Graph analytics and semantic metadata management
- The only engine passing both LDBC Semantic Publishing and Social Network Benchmarks!



Next Generation of AI-Enabled Applications

ontotext
THE GRAPHDB COMPANY

Impact Radar for 2024



Source: Gartner (January 2024)
791104

Knowledge Graphs

Range: 0 to 1 Year
The range for knowledge graphs is Now (down from one to three years in 2022), as KG adoption has rapidly accelerated in conjunction with the growing use of AI, generally, and large language models (LLMs), specifically. GenAI models are being used in conjunction with KGs to deliver trusted and verified facts to their outputs, as well as provide rules to contain the model.

Emerging Tech Impact Radar: Generative AI
16 November 2023

We Enable LLMs to Serve Your Data!

ontotext
THE GRAPHDB COMPANY

Don’t ask: **What graphs can do for LLMs?**

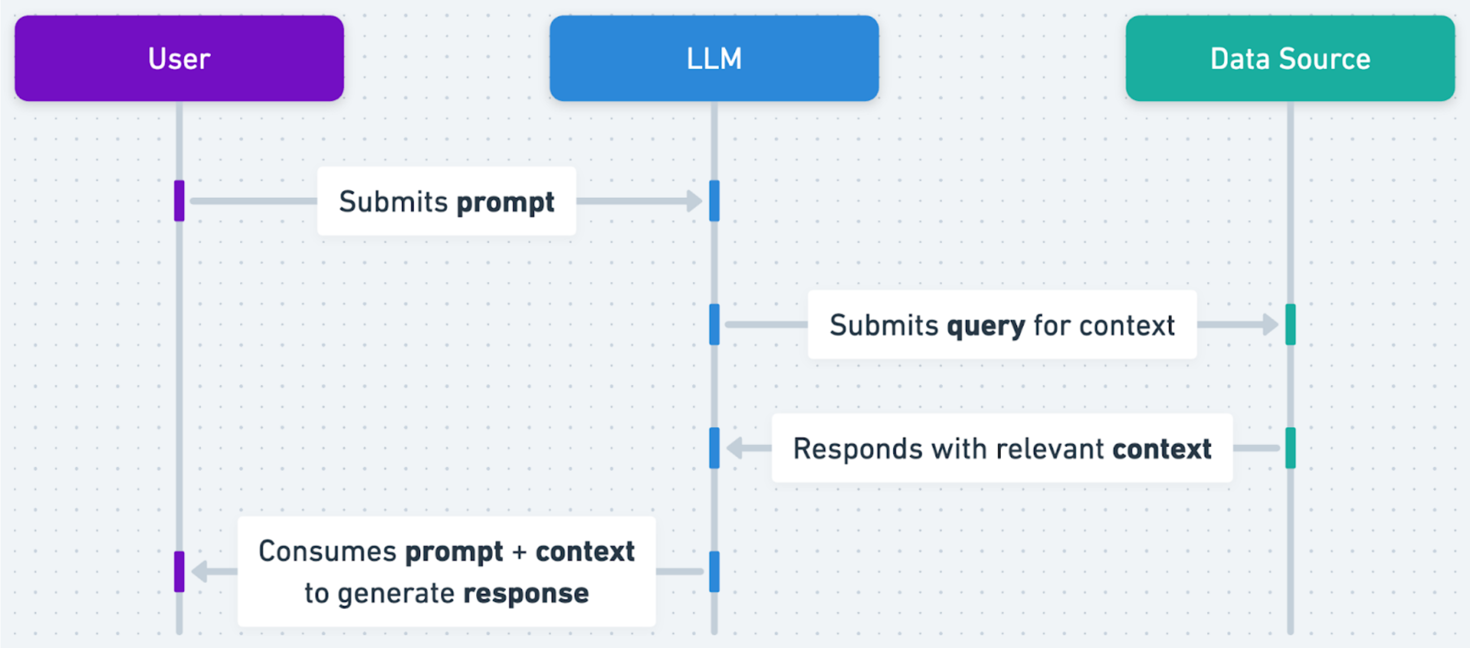
- ✓ **Graph as a fact-checker? No!**
  Validation will become even more expensive than the generation!
- ✓ **Graph as a Teacher? Yes!!!**
  KGs are best for fine-tuning (educating) LLMs and assessing (examining) their capabilities

Ask: **What AI can do to help you use and manage your data?**

- ✓ **Improving discoverability**
- ✓ **Enriching your databases**
- ✓ **Reducing the DM overhead**
- ✓ **Linking proprietary data and domain knowledge**

RAG and Graph RAG

ontotext



Open AI RAG definition

1. **RAG is not only about document chunks**
2. **Vector databases are not new**
3. **Generative AI is not just about RAG**

Let AI Help Data Management

ontotext

1. **Talk to your data to get instant insights**
2. **Turn text into structured data,**
   extract new facts on the fly
3. **Accelerate KG development**
4. **Categorize documents to power KM,**
   e.g. precision search and recommendation
5. **Infer and predict relationships**

Graph RAG Varieties

ontotext

What do we offer?

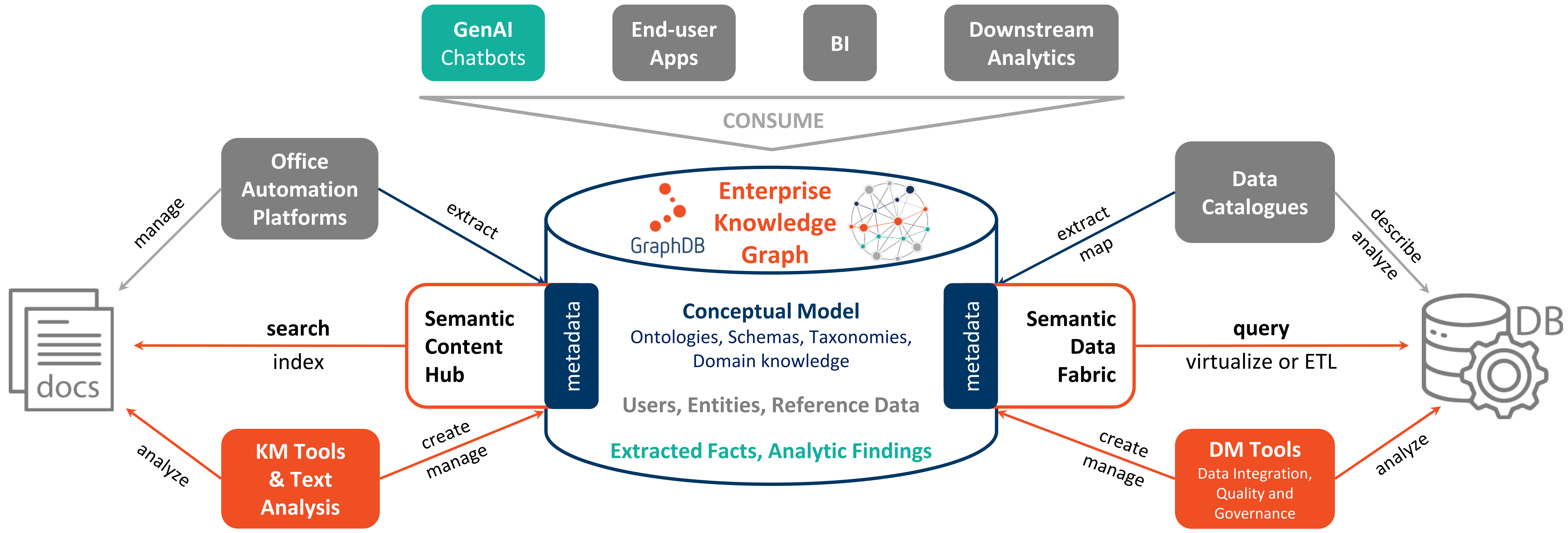
GraphDB and Metadata Studio enable all RAG varieties and combinations

Graph RAG Varieties	Semantic Metadata	Domain Knowledge	Factual Data
Vanilla “chunky” RAG			
Type 1: **Graph as a Metadata Store**	+	–	–
Type 2: **Graph as an Expert**	+	+	–
Type 3: **Graph as a Database**	+	+	+

Read more: [What’s Graph RAG?]

Semantic Layer = Content Hub + Data Fabric

ontotext



Semantic Layer = (Content Hub + Data Fabric) * (Conceptual Model) ^ GenAI

Graph RAG Varieties

ontotext

What do we offer?

GraphDB and Metadata Studio enable all RAG varieties and combinations

Graph RAG Varieties	Semantic Metadata	Domain Knowledge	Factual Data	WHAT GRAPH?
Type 1: **Graph as a Metadata Store**	+	–	–	**Content Hub**
Type 2: **Graph as an Expert**	+	+	–	**Domain Knowledge**
Type 3: **Graph as a Database**	+	+	+	**Data Fabric**

Read more: **What's Graph RAG?**

KG+AI Solutions Scoping Guide

ontotext

What is the high-level task: Q/A, Data Integration, Graph Completion?

This determines the **technical tasks**: **NLQ, RAG, EL, IE, Document Tagging, Reconciliation, Link Prediction**

Non-Functional Requirements:

- **How critical is accuracy?**
- **Some non-determinism is OK?**
- **Language:** English, Spanish, Multilingual, …
- **Domain:** General, LSC, FSI, …
- **Can we reuse a public ontology:** Yes/No
- **Design pattern**: Data Fabric, Content Hub, …

Making an efficient solution requires:

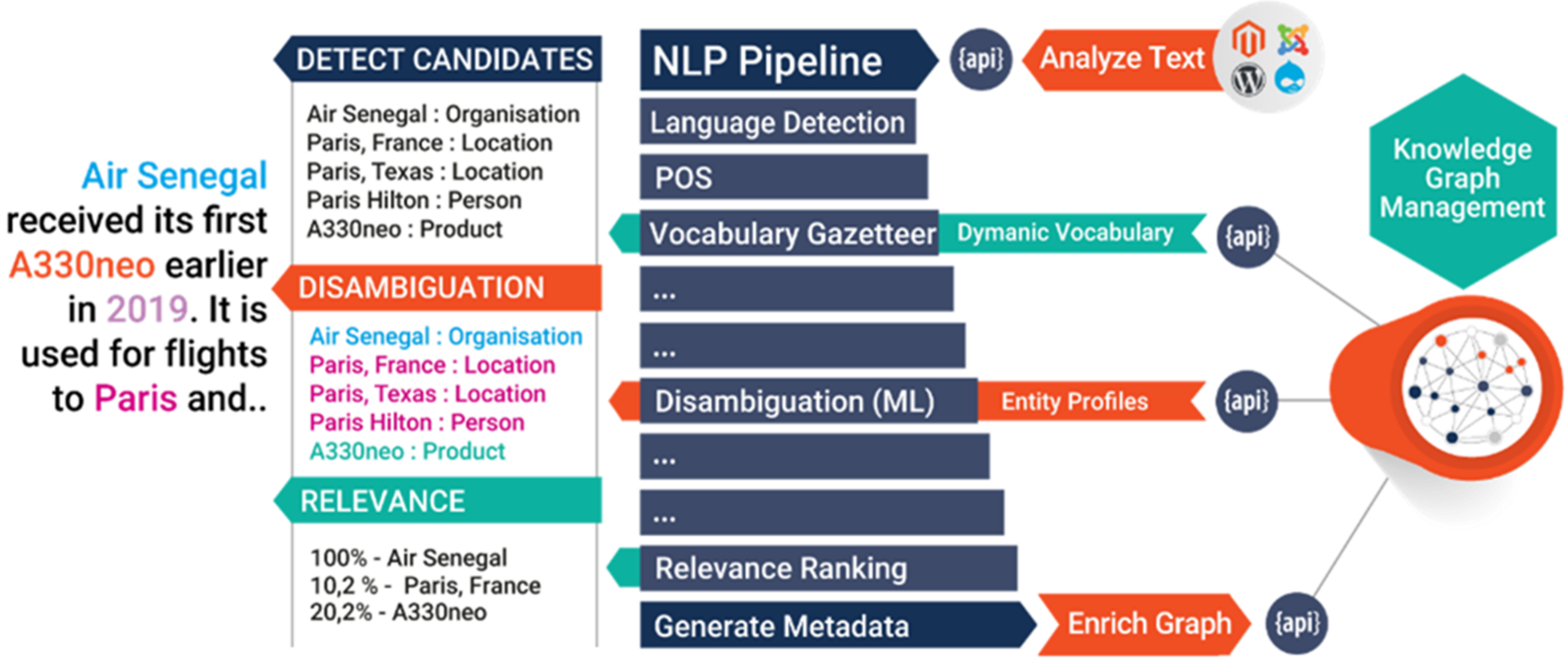
1. **End-to-end graph platform**
2. **Extensive AI toolset**
3. **Skilled KG architect**

Entity Linking: The Task That LLMs Can't Do

ontotext

What is it? – Associate named entity references in text with concrete entity identifiers

Why? – Improve the performance of NLQ, RAG, and information extraction



Entity Linking: The Task That LLMs Can't Do

What do we offer?

Multiple state-of-the-art options

Read more: [What is Entity Linking?]

Entity Linking System	Reference Data or Graph	Model / Technology	Multi-lingual	Accuracy (AIDA)	Speed
Ontotext CEEL	**Wikidata** (38M concepts recognized out of 100M)	RefinED/ RoBERTa	**EN**	76%	2 sec./doc. CPU
BELA	**Wikipedia** (7M concepts)	RefinED	**M/L**	74%	4 sec./doc. GPU
Google NLP	**Wikidata**		**EN**	72%	Slower
Facebook GENRE	**Wikidata**		**EN**	61%	Slower
Ontotext GEL	***General purpose EL*** *(e.g. custom taxonomy + CrunchBase)*	*ChatGPT + SBERT, …*	***EN***	*Depends*	*Depends*
Ontotext Medical MEL	**SNOMED** medical ontology	ClinicalBERT+ KRISSBERT	**M/L**	[#7 of 500 at the 2024 SNOMED Challenge]	

AI Offerings: Product Features, Models & Services

Offering	AI Features and Capabilities	What it is Good For
GraphDB	**ChatGPT Retrieval Connector** with customizable vector DB indices	Zero-code RAG implementation fetching relevant data from a KG
	Talk To Your Graph chat UI in the Workbench	RAG and NLQ experimentation and pilots
	GraphDB NLQ Chain in LangChain (Python)	Queries from the LangChain framework
	SPARQL functions for ChatGPT integration	Summarization of query results, data transformation, information extraction, …
	Similarity Search via word/graph-embedding	Retrieve similar entities and documents
Metadata Studio	Environment for the development of NLP pipelines and metadata management, including tagging and extraction	IDE for data scientists, data engineers, and knowledge workers
CEEL	English entity linking for Wikidata	Semantic tagging, information extraction
Reconciliator	Strings-to-things reconciliation service Pre-trained for Wikidata	Map string from tables to Wikidata IDs

Combine Graphs and AI with Ontotext

ontotext

1. **Improve GenAI performance via quality data and domain knowledge**
   - ✓ Unified view towards diverse data, governance, contract on meaning and data quality via validation
   - ✓ We provide context to your data, by linking to public and 3rd party data
   - ✓ Logical reasoning complements stochastic inference models
2. **GraphDB: The best database engine to serve NLQ and Graph RAG**
   - ✓ Only graph engine **audited** to handle efficiently both **graph analytics** and **metadata management** workloads
   - ✓ Combination of different search/IR techniques; **always up-to-date vector database and FTS indices**
   - ✓ Efficient handling and **predictable performance** of diverse workloads
3. **Text analysis tooling and AI models, complementary to the LLMs**
   - ✓ **Tooling for text analysis**: Human in the loop, quality control, comparison and combination of different models
   - ✓ **Entity linking models:** Let you accurately combine LLMs and KGs; state-of-the-art performance proven with benchmarks
   - ✓ **Relate text to structured data:** Enable NLQ and improve the performance of LLMs, without extensive finetuning

Graph+AI Opinion Maker

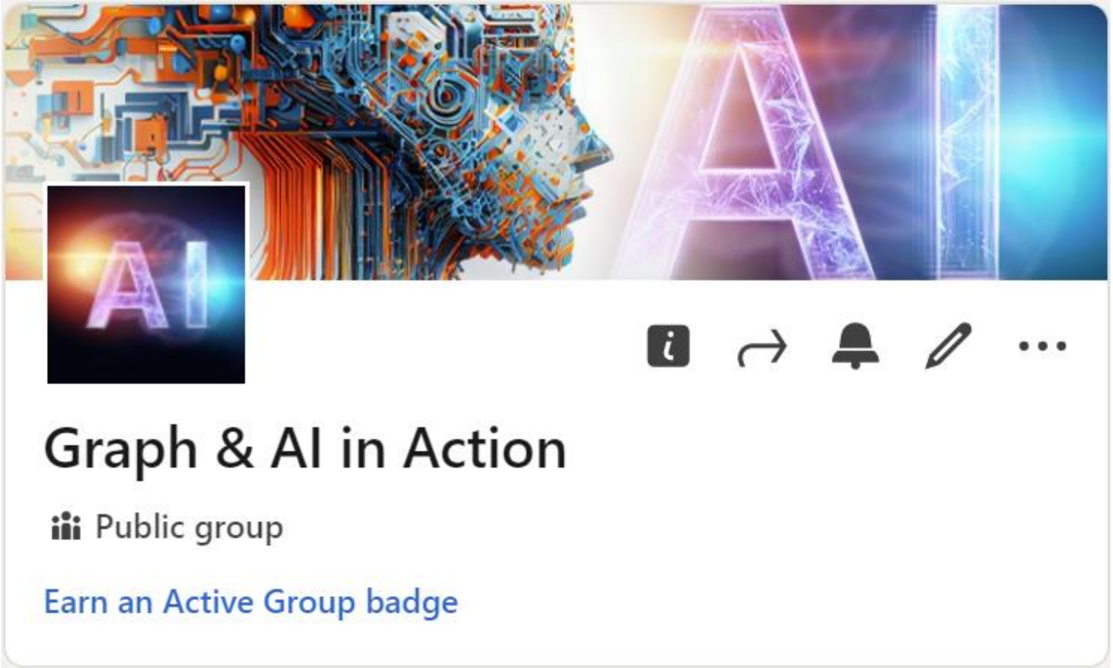
ontotext

We are on 1st page in Google for:

- ✓ Knowledge graphs (#3)
- ✓ Linked data (#1)
- ✓ Semantic annotation (#1)
- ✓ SPARQL (#1)
- ✓ Graph RAG (#1)
- ✓ Entity linking (#3)
- ✓ NLQ (#3)

Join the Graph & AI in Action LinkedIn Group!

And promote your KG+AI story



SMOOTH DATA INTEGRATION

We make AI help you use and manage your data

ontotext

The "NASA RAG Problem"

ontotext

The task: Find a document mentioning NASA when searching for "Government Agency"

Demo Context: Ask the OTKG chat about "U.S. Gov customers of Ontotext"

The challenges:

1. Naïve RAG doesn't work, because there is no mention of government in the chunk
2. There are too many chunks that mention U.S. Gov. organizations

The Chunky RAG Limitations

Graph RAG Type 2 approach:

1. Use **CEEL Entity Linking** to tag documents with mentions of Wikidata entities
2. Enrich chunks with **entity descriptions** from WD
   - The enriched chunks, mentioning NASA, also mention US Gov. as part of NASA's description
   - We also implemented question enrichment
3. Use **GraphDB's ChatGPT Retrieval Connector** for zero-code RAG

It didn't solve "The NASA problem"

1. There were 100s of chunks relevant to the question
2. We can only pass ~10 to ChatGPT
3. All the 4 chunks mentioning NASA scored beyond position 20, because there were more relevant chunks

The Chunky RAG Axiom:
If the information required to answer the question is spread across more chunks than what we can pass to the LLM, the answer will be incomplete.

Read the full story in "Okay, RAG… We Have a Problem" AI-in-Action blog post

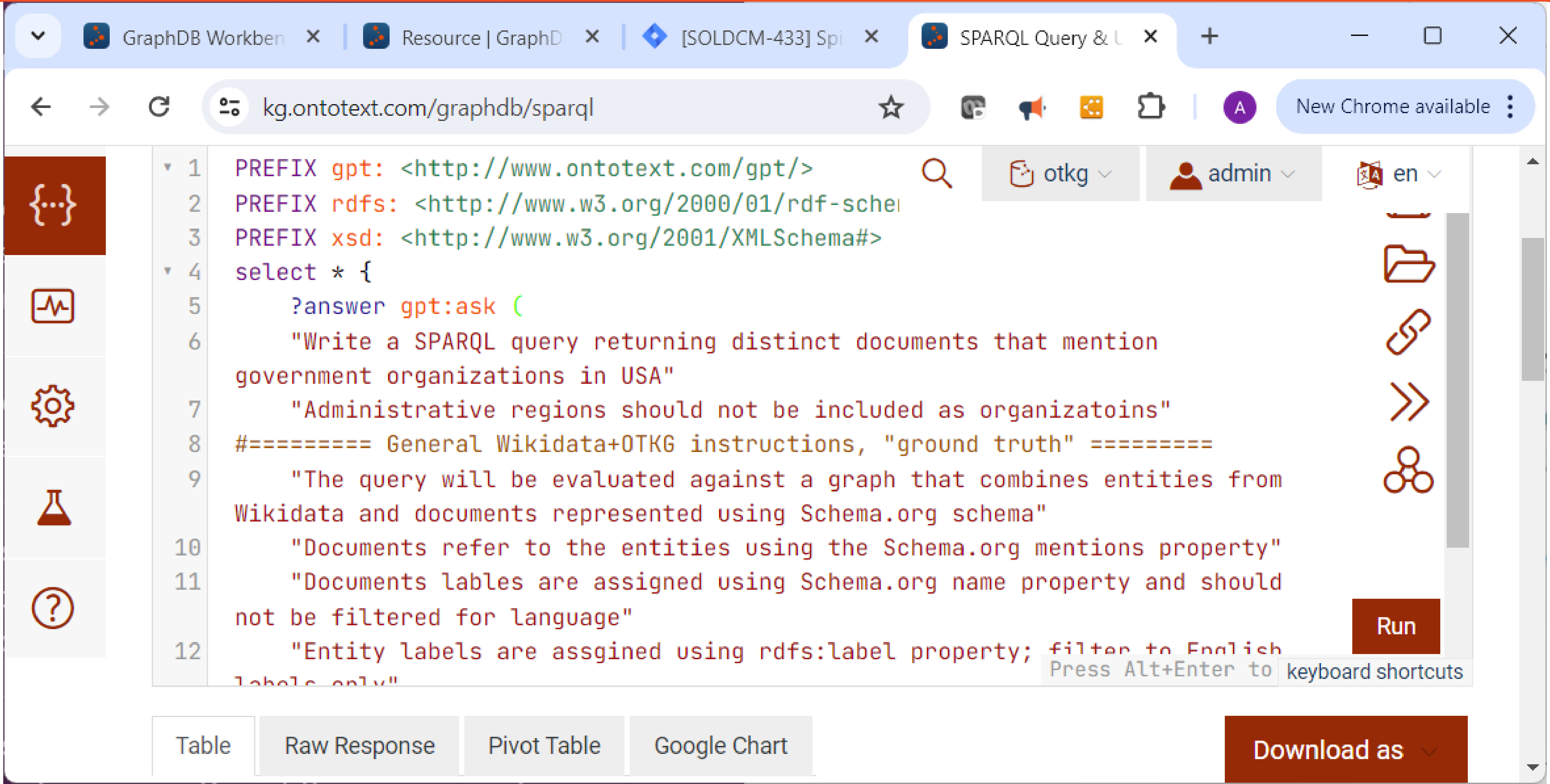
The NLQ Solution: ChatGPT Knows Wikidata!

ontotext

Graph RAG Type 3 – Graph as a Database:

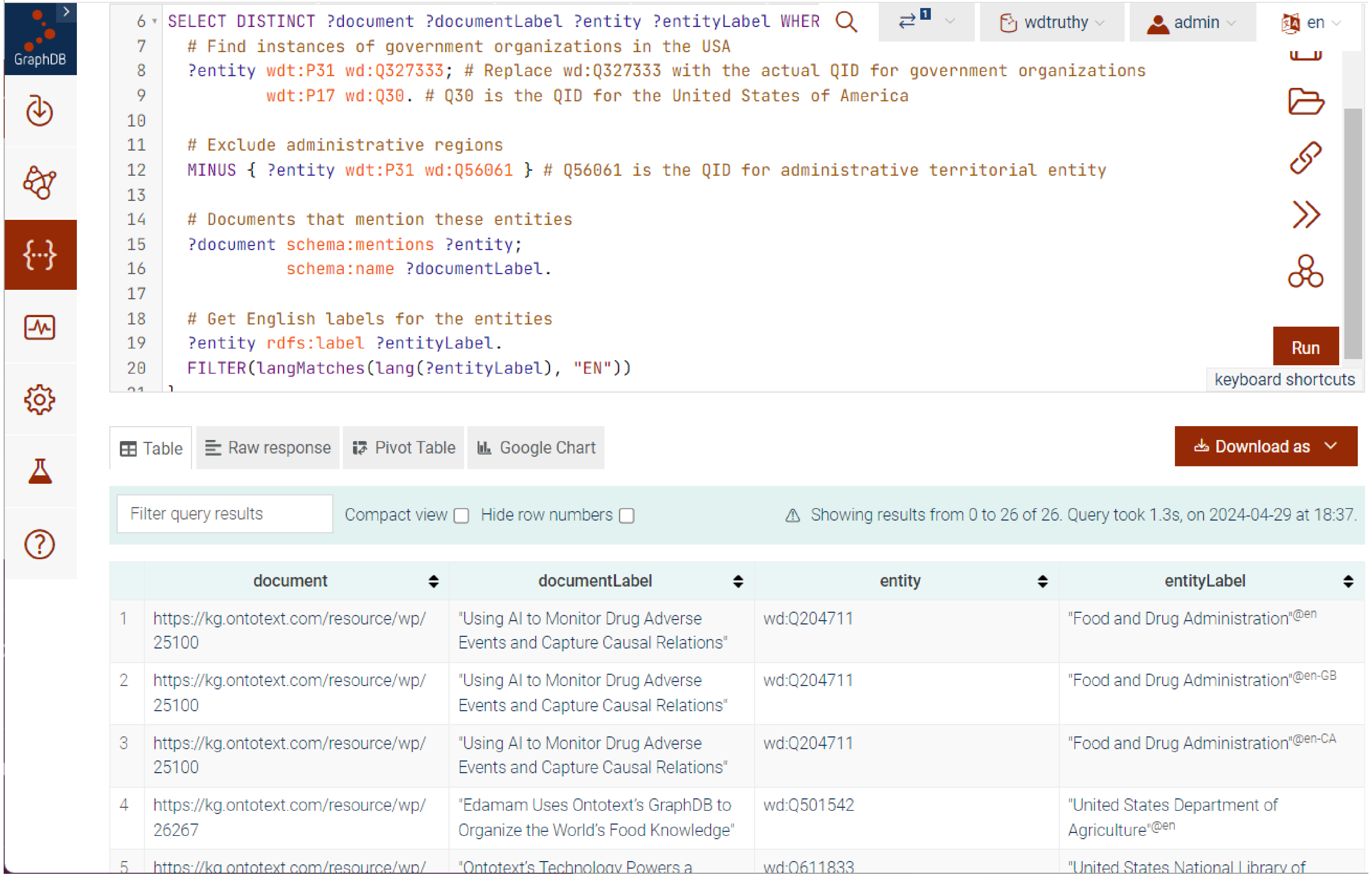
1. Load your documents in **GraphDB**
    - Use **Schema.org** to represent document metadata – it is the most popular schema for this and ChatGPT knows it
2. Use CEEL to **link documents to Wikidata entities**
    - Process your documents with a single SPARQL query and get your graph extended with statements like `<my:document schema:mentions wd:entity>`
3. Ask ChatGPT to **convert NL question to SPARQL**
    - Tell ChatGPT what's in the graph and what schemata are used
4. **Get the relevant documents and entities**
    - Evaluate the SPARQL query
    - Access Wikidata via federation or load your data next to Wikidata replica repo (ask Ontotext for man. service)
5. **Pass these documents and entities** to an LLM to answer the NL question (optional)

The NLQ Solution: ChatGPT Knows Wikidata!



The NLQ Solution: ChatGPT Knows Wikidata!

ontotext



Why does this experiment matter?

ontotext

1. Wikidata is an example of a **general-purpose source of "semantic context"**
2. Every engineer can implement **custom Graph RAG with 5 SPARQL queries**, zero code
3. One can **combine existing ontologies with proprietary data**
4. **We can provide all the tools, AI models and services needed** to:
   - Let your engineers play with the data and develop the basics
   - Make a PoV and demonstrate it to customers or sponsors
   - Control the quality and manage production system