Large Language Models
x
Materials Science Databases

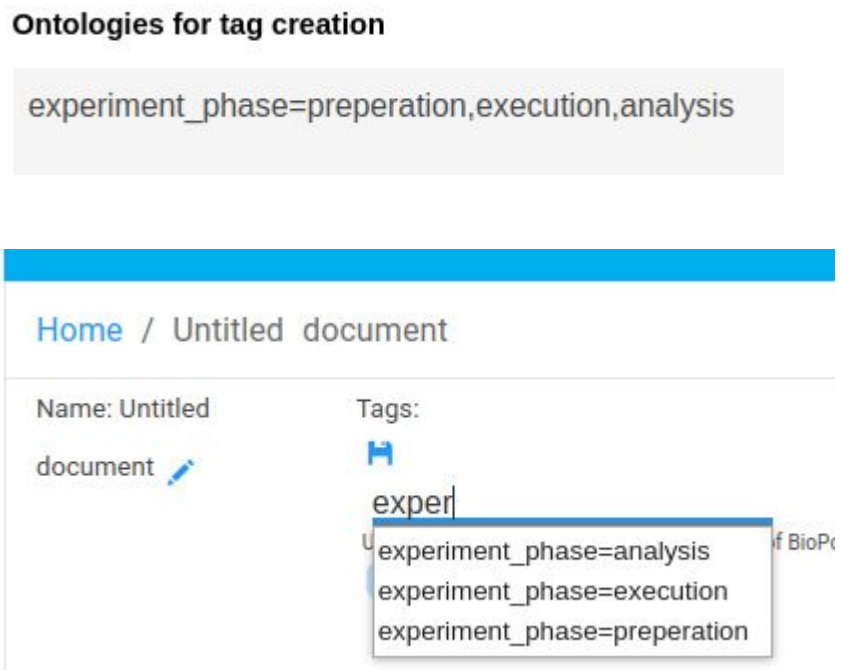
1. RSpace - Example Use Case of LLM
2. Short Theory on Retrieval Augmented Generation
3. Two High-Throughput Experimentation DBs - MPS and MekG
4. Outlook: LLMs x Materials Science Databases

Document-Tags in RSpace using 'Ontology'

RSpace Community Edition:
+ sharing tags in a lab group
+ namespace variables

- manual insertion of tags
- no hierarchical relationships

Idea: Use LLM to retrieve relevant classes from the BattINFO ontology

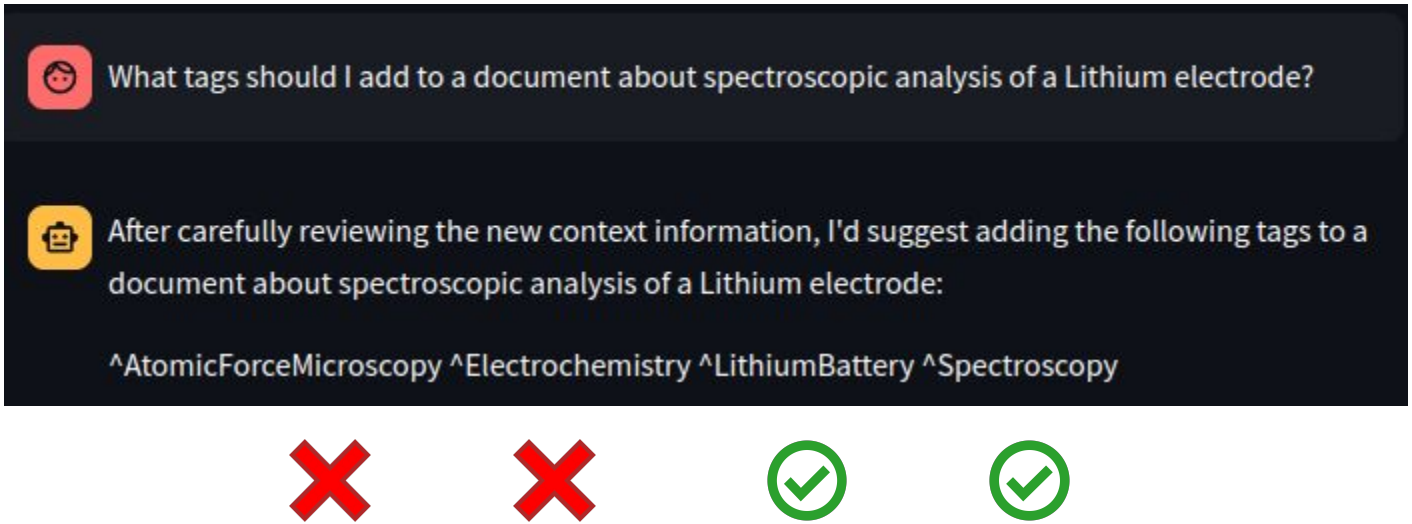


Whats the unit for measurements of the cell potential?

After rewriting the original answer to fit the new context, I found that the unit for measurements of the cell potential is Volt (V) and its related units such as:

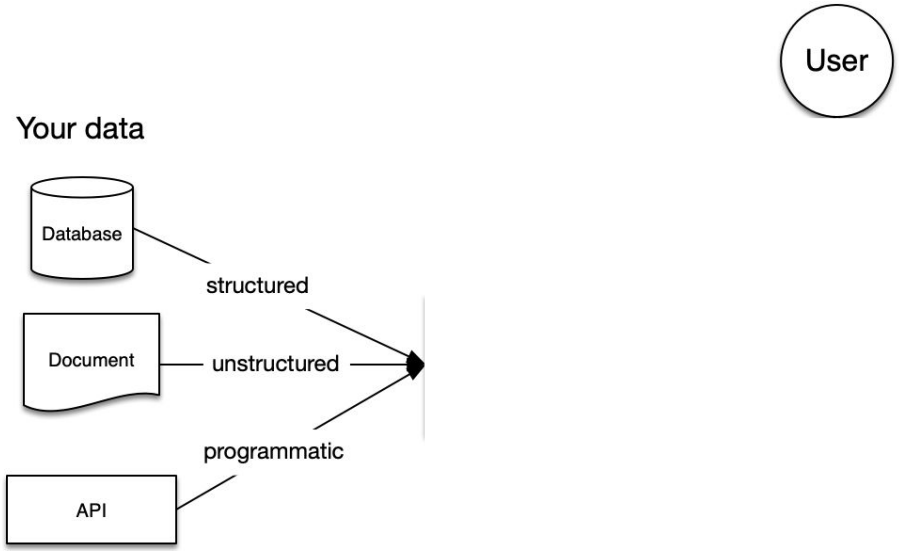
Volt VoltPerCentiMetre VoltPerKelvin VoltPerMetre VoltPerMicroSecond VoltPerMilliMetre VoltPerSecond VoltPerSquareMetre VoltSecondPerMetre

So, my answer is: ^Volt


What tags should I add to a document about spectroscopic analysis of a Lithium electrode?
After carefully reviewing the new context information, I'd suggest adding the following tags to a document about spectroscopic analysis of a Lithium electrode:
^AtomicForceMicroscopy ^Electrochemistry ^LithiumBattery ^Spectroscopy

Retrieval Augmented Generation

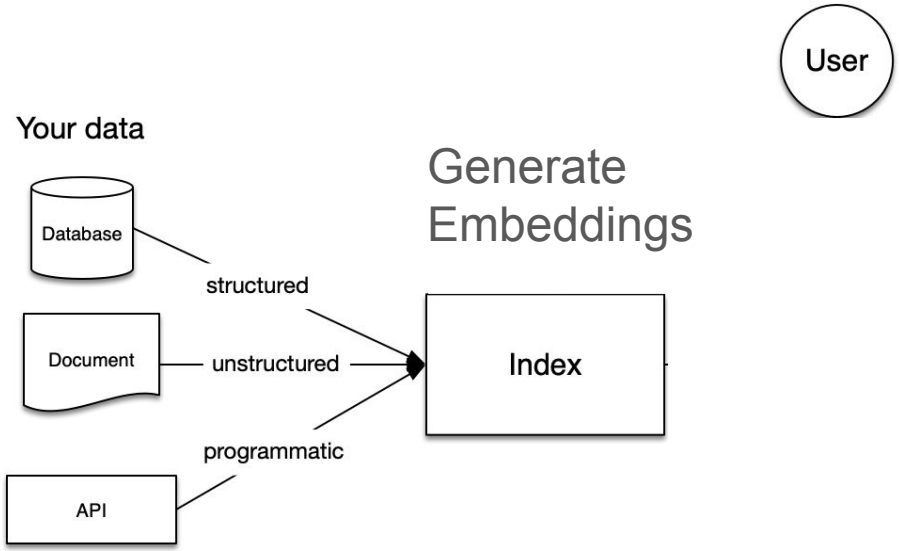
LLMs hallucinate -> RAG provides facts as context when prompting an LLM



https://docs.llamaindex.ai/en/stable/getting_started/concepts/

Retrieval Augmented Generation

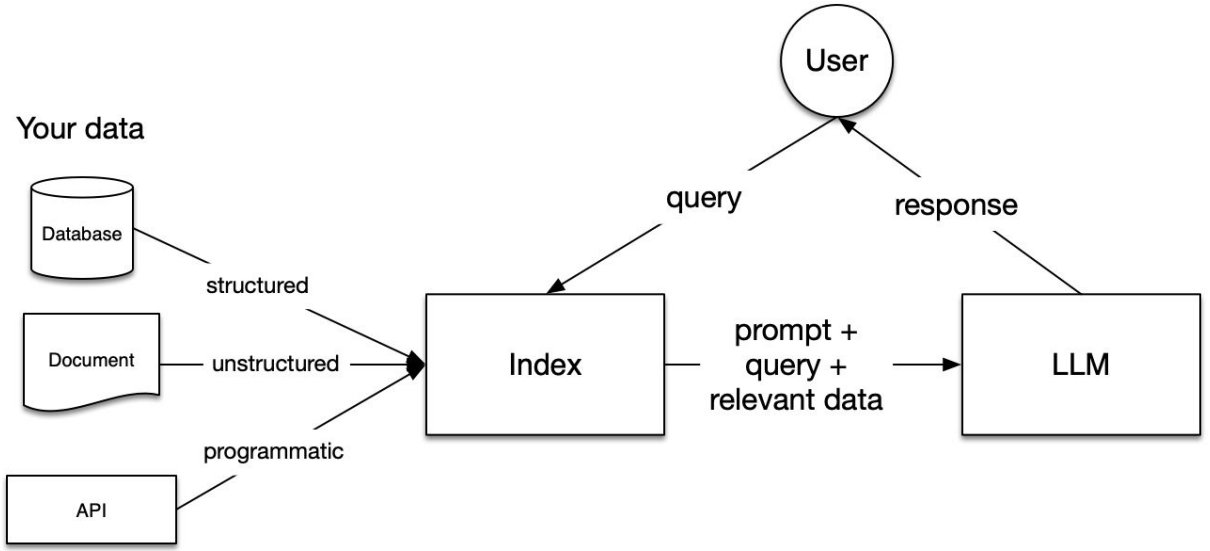
LLMs hallucinate -> RAG provides facts as context when prompting an LLM



https://docs.llamaindex.ai/en/stable/getting_started/concepts/

Retrieval Augmented Generation

LLMs hallucinate -> RAG provides facts as context when prompting an LLM



https://docs.llamaindex.ai/en/stable/getting_started/concepts/

databases from HTE of metal oxide solid state materials

synthesis, characterization and analysis

The materials provenance store

Relational DB

Statt et al. Scientific Data 2023 10:184

The materials experiment knowledge graph

GraphDB

Statt et al. Digital Discovery 2023,2, 909

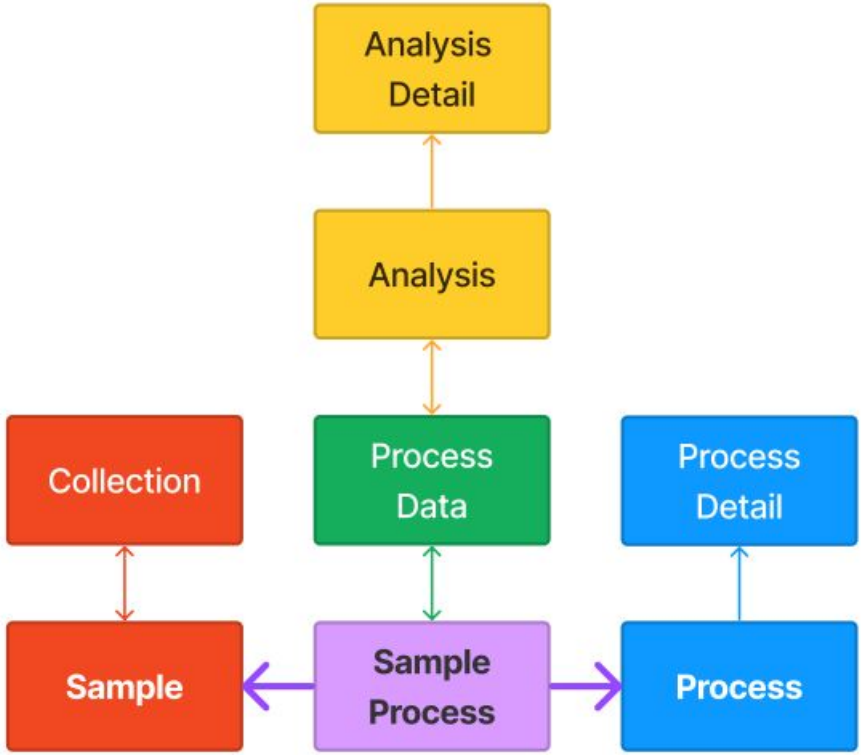
MPS - Relational DB

samples: 11.2M

sample-process: 24M

processes: printing, annealing, electrochemistry, diffraction, spectroscopy etc.

analysis: figure-of-merit

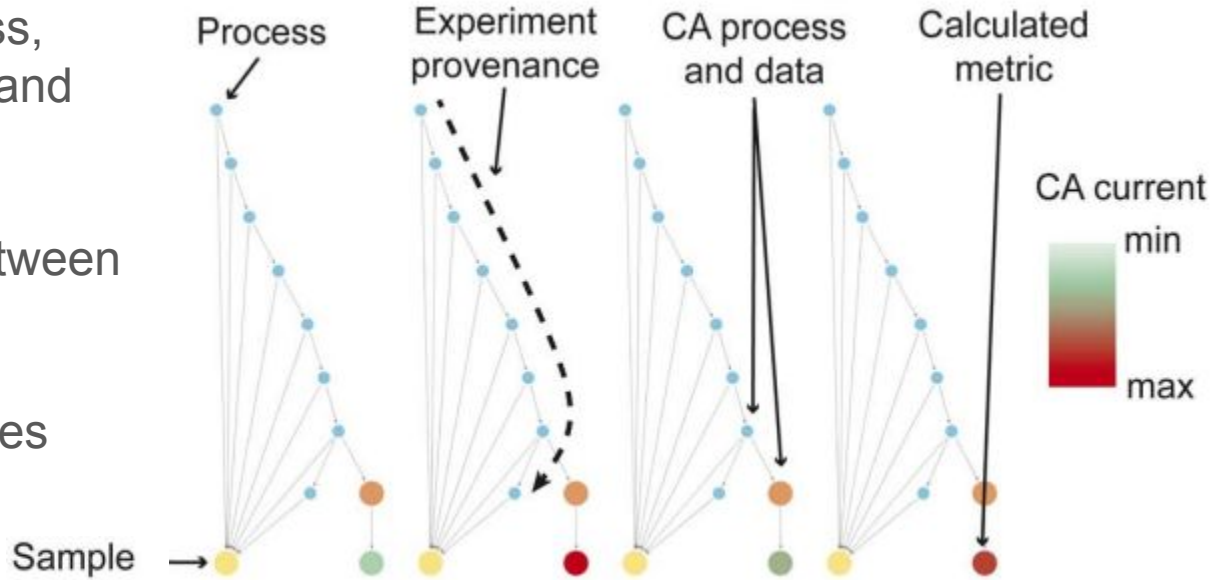


MekG - GraphDB

nodes: sample, process, sample-process, exp. and analysis details

edges: relationship between nodes

52M nodes, 110M edges



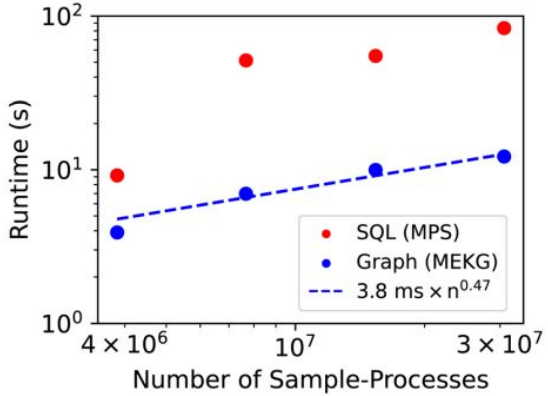
Discussion - Comparison MPS and MekG

preparation time:

(very complex queries)

MekG: ~mins, MPS: ~days

and execution time:



"graph-based queries are sufficiently fast for real-time data exploration"

"graph schema more intuitive than the SQL schema"

Use Case - Automate Design of Experiments

Prior knowledge

"correlation of OER activity in pH 3 and pH 7 electrolytes among metal oxide catalysts"

DB Query

Find catalysts that have been tested at pH 7 but not at pH 3
Find catalysts that have been synthesized but not tested

Experiment

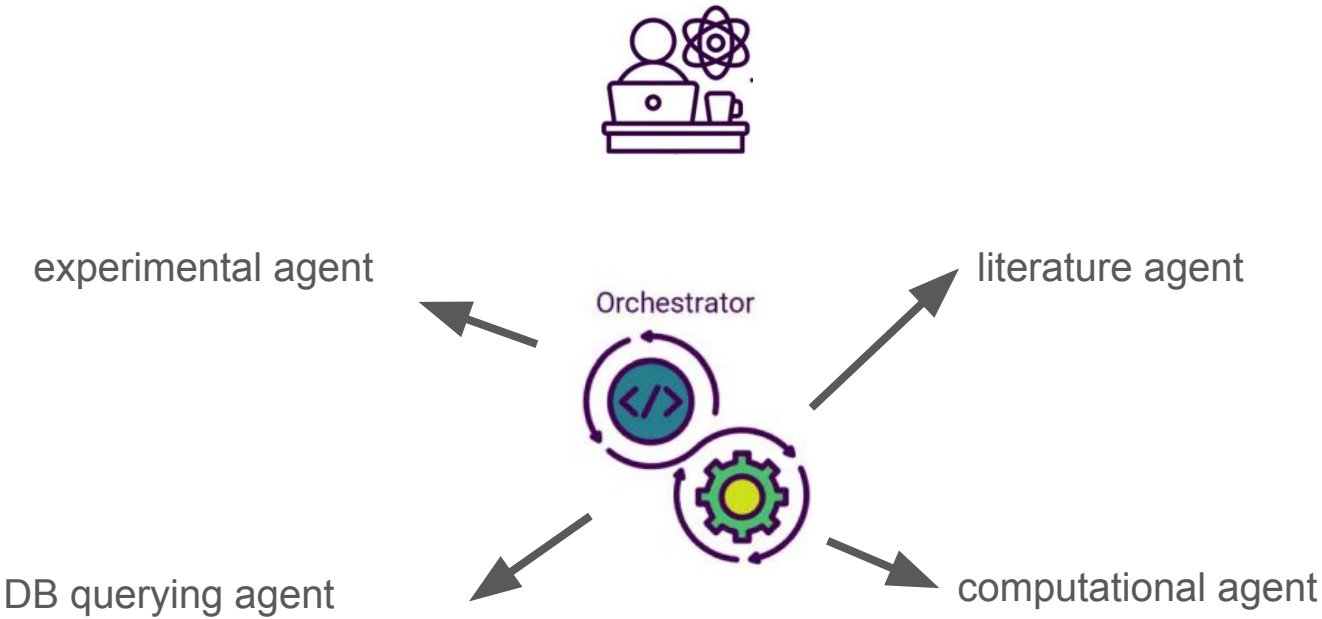
69K new activity measurements proposed

Use Cases of LLMs + Materials Science DB

1. Text2Query - Instead of writing queries in Cypher or SQL, a LLM converts natural language + the schema of the database to create the DB query

2. AI Agents - Instead of the user defining a series of actions to perform, specialized agents (LLMs) prompt each other. User only gives a goal, and the model figure out the steps themselves

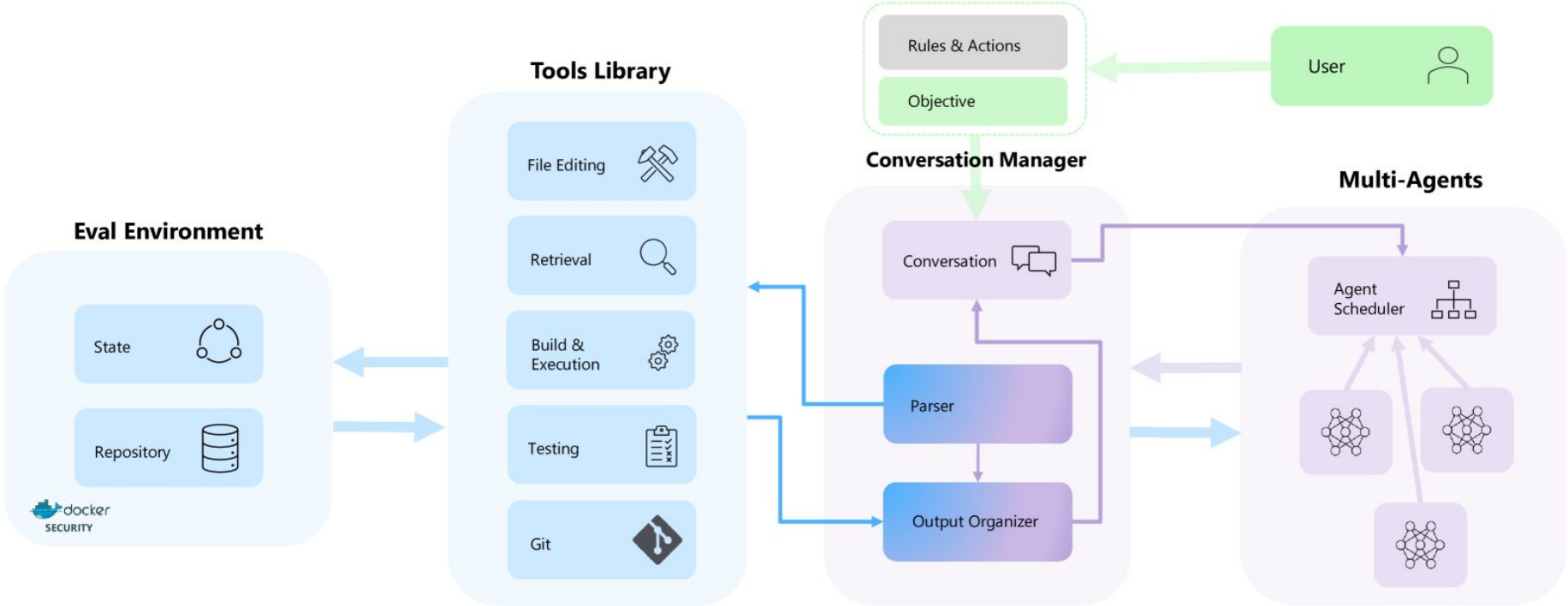
Idea - Design of Experiments by interacting LLM Agents



Backup

Human researchers possess domain expertise combined with intuition from their aggregated prior knowledge, both of which are unrivaled by machine learning to-date.

Auto Dev - Tufano et al. - 2024 - arXiv:2403.08299v1



Agentic AI Systems

Code Generation for Computer Vision: va.landing.ai/chat/

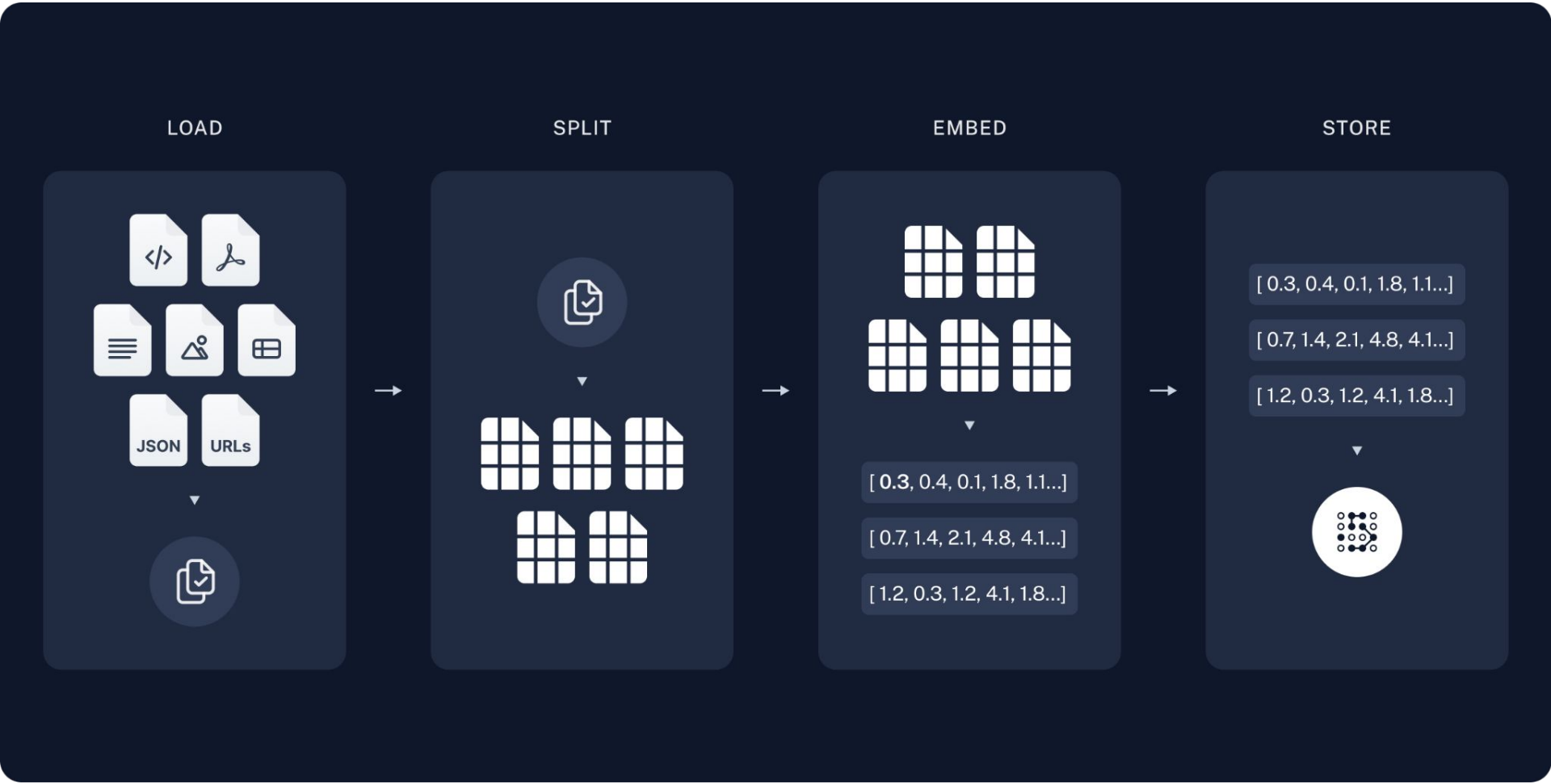
Robotics: Figure AI - https://www.youtube.com/watch?v=Q5MKo7Idsok

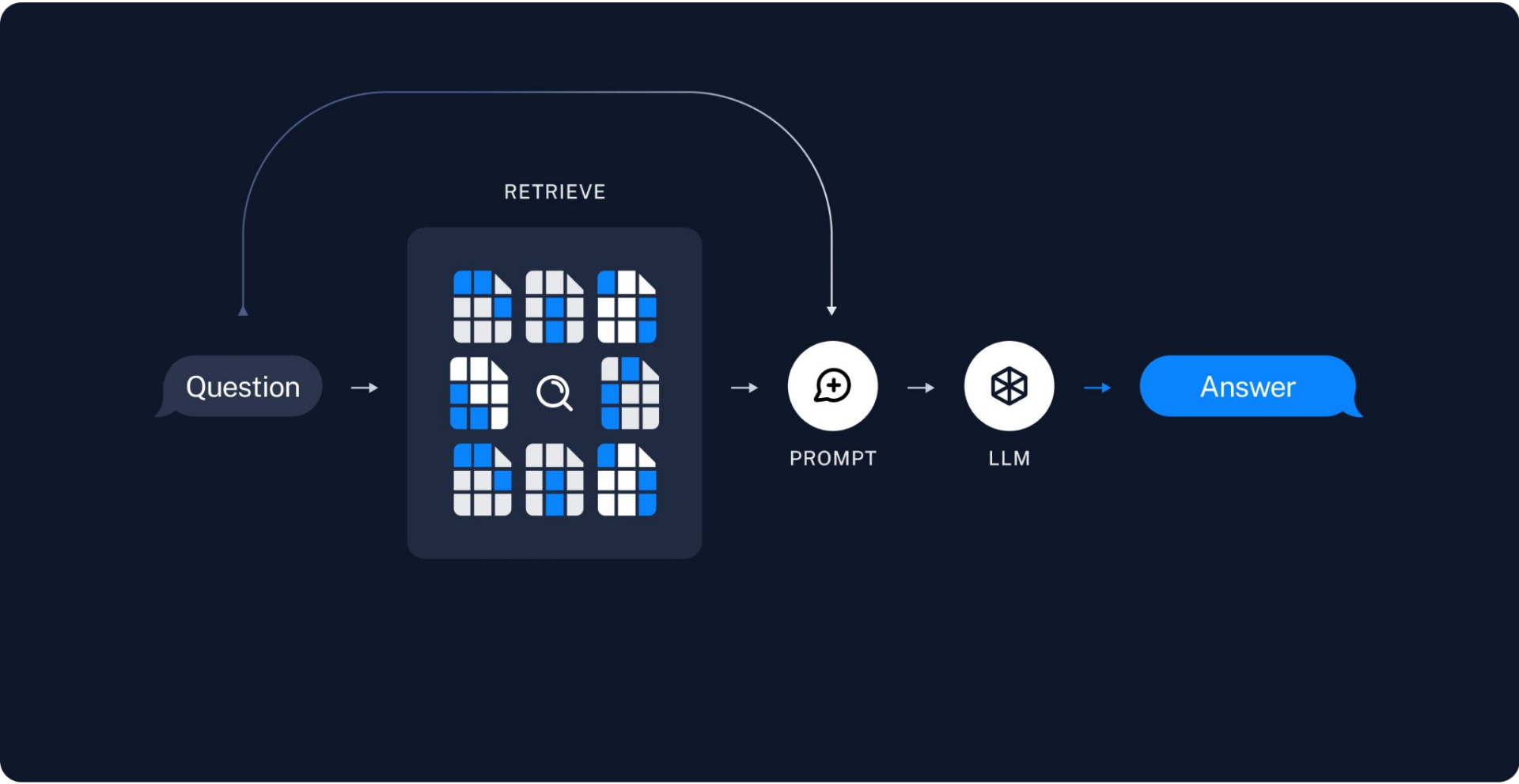
Prompt for BattINFO tagging

```
"Context information about the ontology is below.\n"
"---------------------\n"
"{context_str}\n"
"---------------------\n"

"Given the context information above I want you to think step by step to answer the query in a crisp manner, "
"incase case you don't know the answer say 'I don't know!'.\n"
"Query: {query_str} Output only a list of tags, seperated by a '^': ^CycleLife ^LithiumAirBattery ^R2012."
"Make sure all tags are in the context information above.\n"
"Answer: "
```

RAG - Pipeline


LOAD
SPLIT
EMBED
STORE
JSON
URLs
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]
[0.3, 0.4, 0.1, 1.8, 1.1...]
[0.7, 1.4, 2.1, 4.8, 4.1...]
[1.2, 0.3, 1.2, 4.1, 1.8...]


RETRIEVE
Question
PROMPT
LLM
Answer