AI

Summer of Code

MongoDB®

Developer Data Platform

MongoDB Atlas, MongoDB Charts, Atlas Edge Server(Public Preview), Triggers, Stream Processing

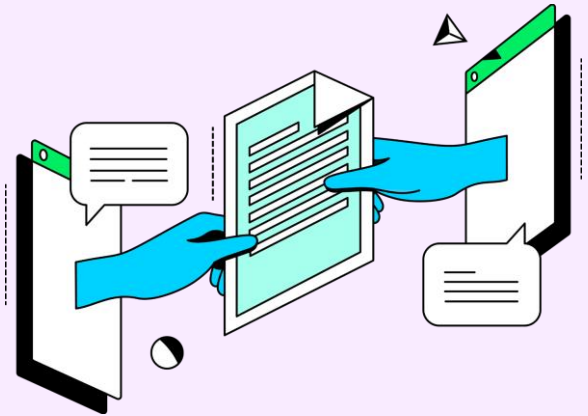
Document Model

JSON/BSON data model, Powerful, Universal, Flexible, Familiar

AI/LLM Applications

Vector Database for RAG applications, Memory Provider for Agentic Systems

```
{
  "_id": ObjectId("5f8a7b2b9d3b2a1b1c1d1e1f"),
  "name": {
    "first": "Jane",
    "last": "Doe"
  },
  "position": "Senior Data Scientist",
  "department": "AI Research",
  "skills": ["machine learning", "python", "data visualization"],
  "performance": {
    "latest_score": 4.8,
    "review_date": ISODate("2023-06-30")
  },
  "hire_date": ISODate("2019-03-15"),
  "embedding": [0.123, 0.456, 0.789, ..., 0.321]
}
```



MongoDB

LLM Application Landscape and Everything In Between

Form Factor Evolution

LLM Powered Chatbots

Text box that fits up to three lines of copy. Do not resize font or box.

RAG Chatbots

Text box that fits up to three lines of copy. Do not resize font or box.

AI Agents

Text box that fits up to three lines of copy. Do not resize font or box.

Agentic Systems

Text box that fits up to three lines of copy. Do not resize font or box.

BUILDING BLOCKS OF AI INFRASTRUCTURES AND APPLICATIONS

{ AI Stack }

- What is the AI Stack
- High Level view of the AI Stack
- Low Level view of the AI Stack
- Components of the AI Stack

The AI Stack

What is the AI Stack?

The AI stack **combines integrated tools, libraries, and solutions to create applications with generative AI capabilities**, such as image and text generation. The components of the AI stack include programming languages, model providers, large language model (LLM) frameworks, vector databases, operational databases, monitoring and evaluation tools, and deployment solutions.

AI STACK

High Level View

Application

Tooling (Data, Infra)

Compute (GPU and LLMs)

AI STACK

Low Level View

- Programming Language
- Model Provider
- LLM Orchestrators and Frameworks
- Operational and Vector Database
- Monitoring and Observability
- Deployment

GROUNDING LLMS WITH DOMAIN SPECIFIC OR PROPRIETARY DATA

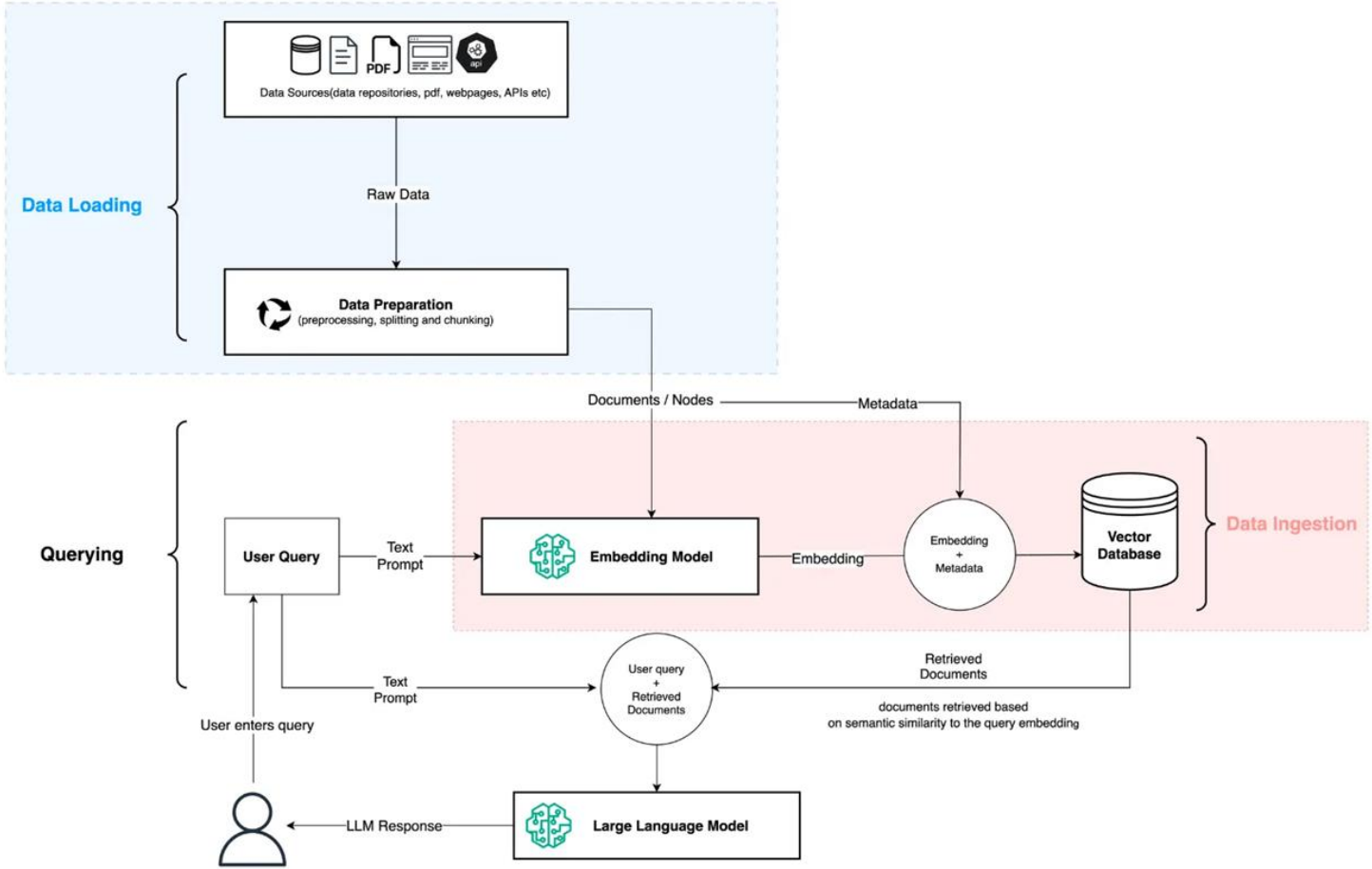
{ RAG }

- What is RAG?
- How does RAG work?
- Benefits of RAG

Retrieval Augmented Generation

What is RAG?

Retrieval-augmented generation (RAG) is a system design pattern that leverages information retrieval techniques and generative AI models to **provide accurate and relevant responses to user queries by retrieving relevant data to supplement user queries** with additional context, combined as input to LLMs.


Data Loading
PDF
api
Data Sources(data repositories, pdf, webpages, APIs etc)
Raw Data
Data Preparation
(preprocessing, splitting and chunking)
Documents / Nodes
Metadata
Querying
User Query
Text
Prompt
Embedding Model
Embedding
Embedding
+
Metadata
Vector
Database
Data Ingestion
User query
+
Retrieved
Documents
Text
Prompt
Retrieved
Documents
documents retrieved based
on semantic similarity to the query embedding
User enters query
LLM Response
Large Language Model

MONITORING AND MEASURING PERFORMANCE OF INTELLIGENT SYSTEMS

{ LLM Evaluation }

- What is LLM Evaluation?
- Why Evaluate LLMs
- Categories of LLM Evaluation
- Metrics for LLM Evaluation

LLM Evaluation

What is LLM Evaluation?

LLM evaluation, also referred to as ‘LLM Eval,’ is the **systematic process of formulating a profile of foundation models or their derived fine-tuned variants** to **understand** and capture their **performance** on certain specialized or general-purpose tasks, **reliability** in certain conditions, **effectiveness** in particular use cases, and many other evaluative measurement criteria that help in gaining an overview of a model’s overall ability.

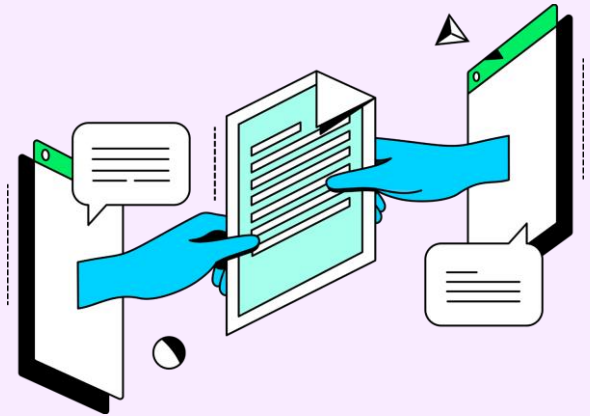
LLM Evaluation is not *LLM Evaluation*

LLM EVALUATION

LLM System Evaluation

Model Evaluation

Why Even Evaluate?



Build PRIME Applications

P	R	I	M	E
Performance Profiling	Risk and Responsibility	Iterative Enhancement	Model Benchmarking	Ethical Safeguarding

AI

Summer of Code