Intro to AI Agents

Yujian Tang | Zilliz

Speaker

Yujian Tang

Senior Developer Advocate, Zilliz

yujian@zilliz.com
https://www.linkedin.com/in/yujiantang
https://www.twitter.com/yujian_tang

zilliz

CONTENTS

zilliz

01 Project Overview

Tech Stack

Milvus

OpenAI

LlamaIndex

zilliz
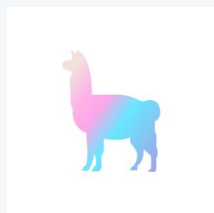
RAG AI Agent Architecture

02

RAG

zilliz

RAG Architecture

zilliz
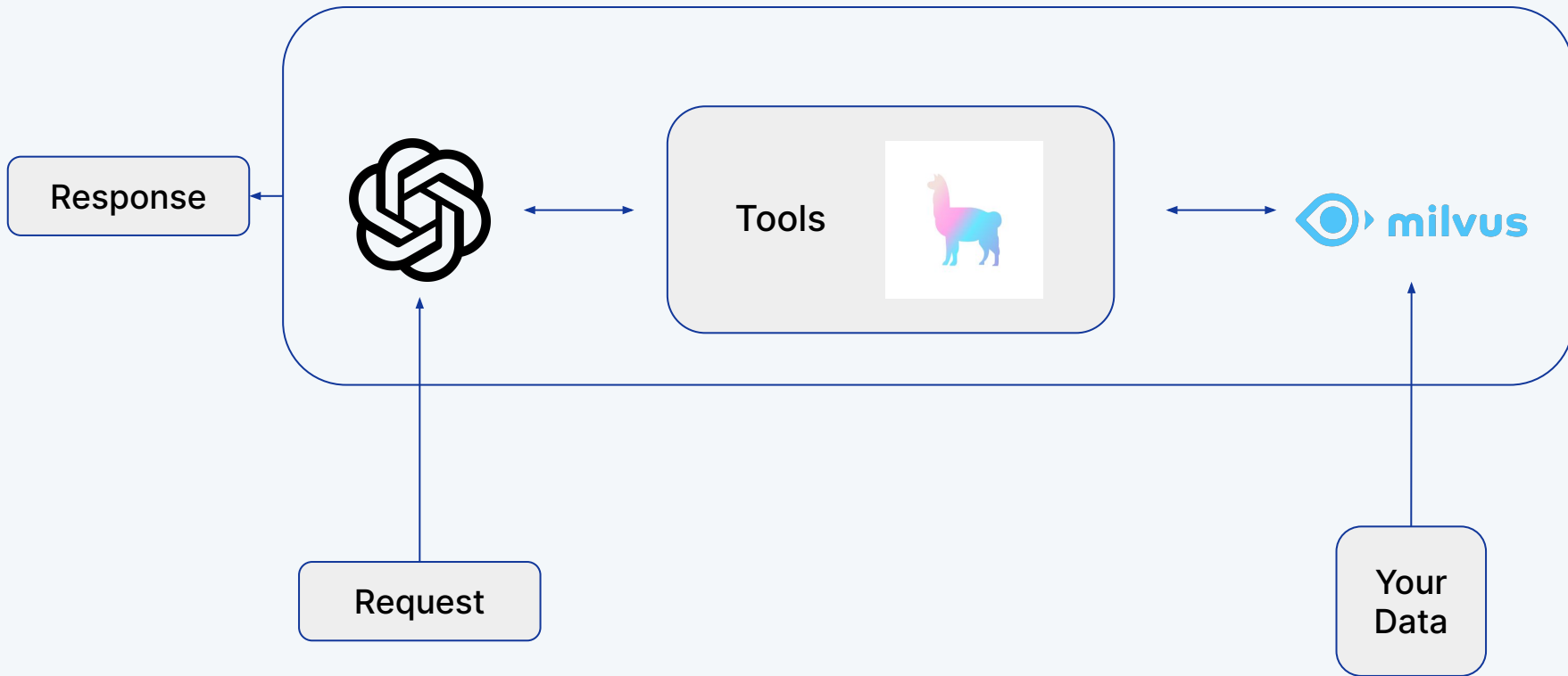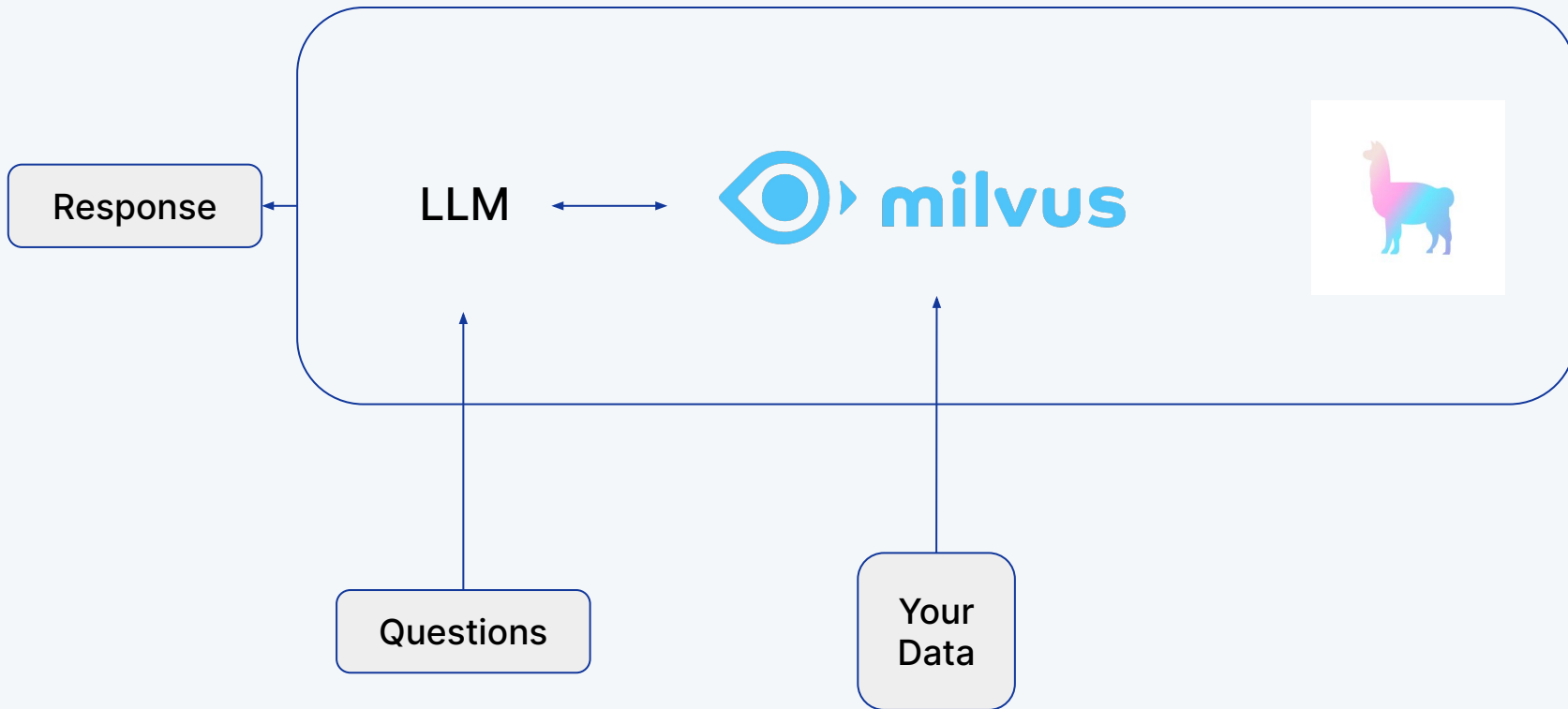
04 AI Agents

AI Agent Architecture

LLM → Tools → LLM

Request → LLM

LLM → Response

zilliz

04

LlamaIndex

zilliz
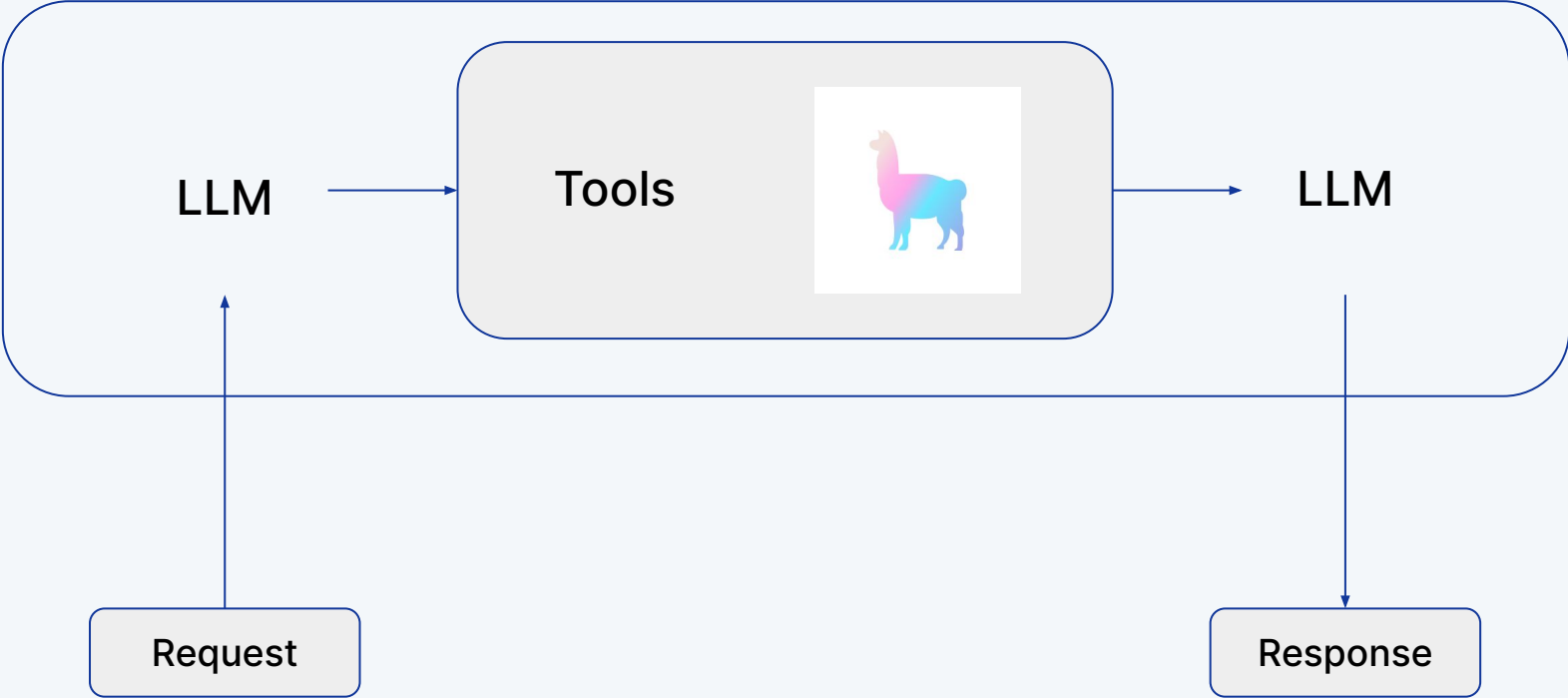
What is LlamaIndex?

- Framework for building LLM Applications
- Focus on retrieving data and integrating with LLMs
- Integrations with most popular tools

zilliz

05 Milvus

What is Milvus?

- Distributed vector database optimized for large scale use cases
- Highly customizable indexes and distance metrics
- Enterprise ready features

zilliz

What Does Vector Data Look Like?

"id": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-9275854

"embedding": [-0.042092223,-0.0154002765,-0.014588429,-0.031147376,0.03801204,0.013369046,0

"date": "2023-06-01"

"paragraph": "We define an anomaly as follows:"

"reading_time": "11"

"subtitle": "A guide for the implementation of an anomaly..."

"publication": "Towards Data Science"

"responses": "1"

"article_url": "https://towardsdatascience.com/detection-of-credit-card-fraud-with-an-autoencoder-

"title": "Detection of Credit Card Fraud with an Autoencoder"

"claps": "229"

↑ Hide 6 fields   🔍 Vector search

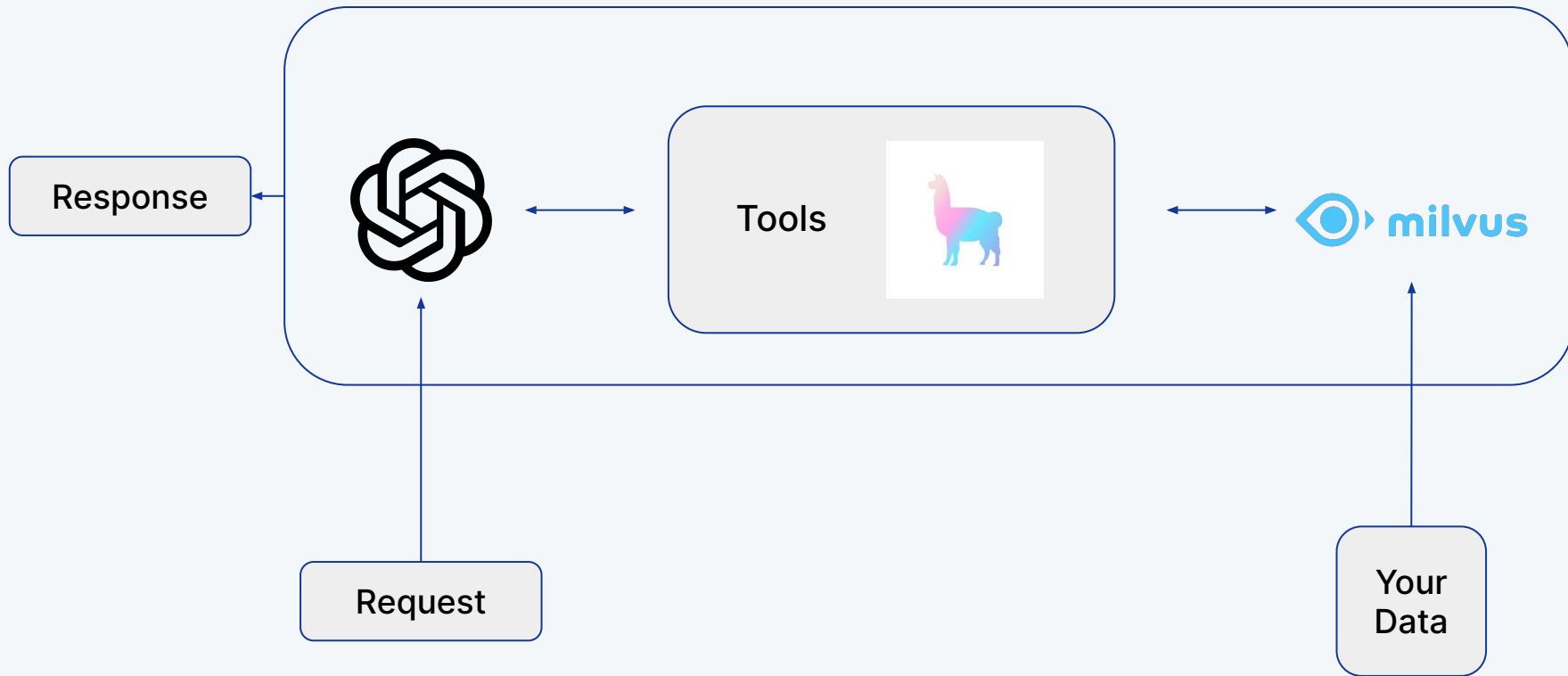✳ zilliz

High-level overview of Milvus' Architecture
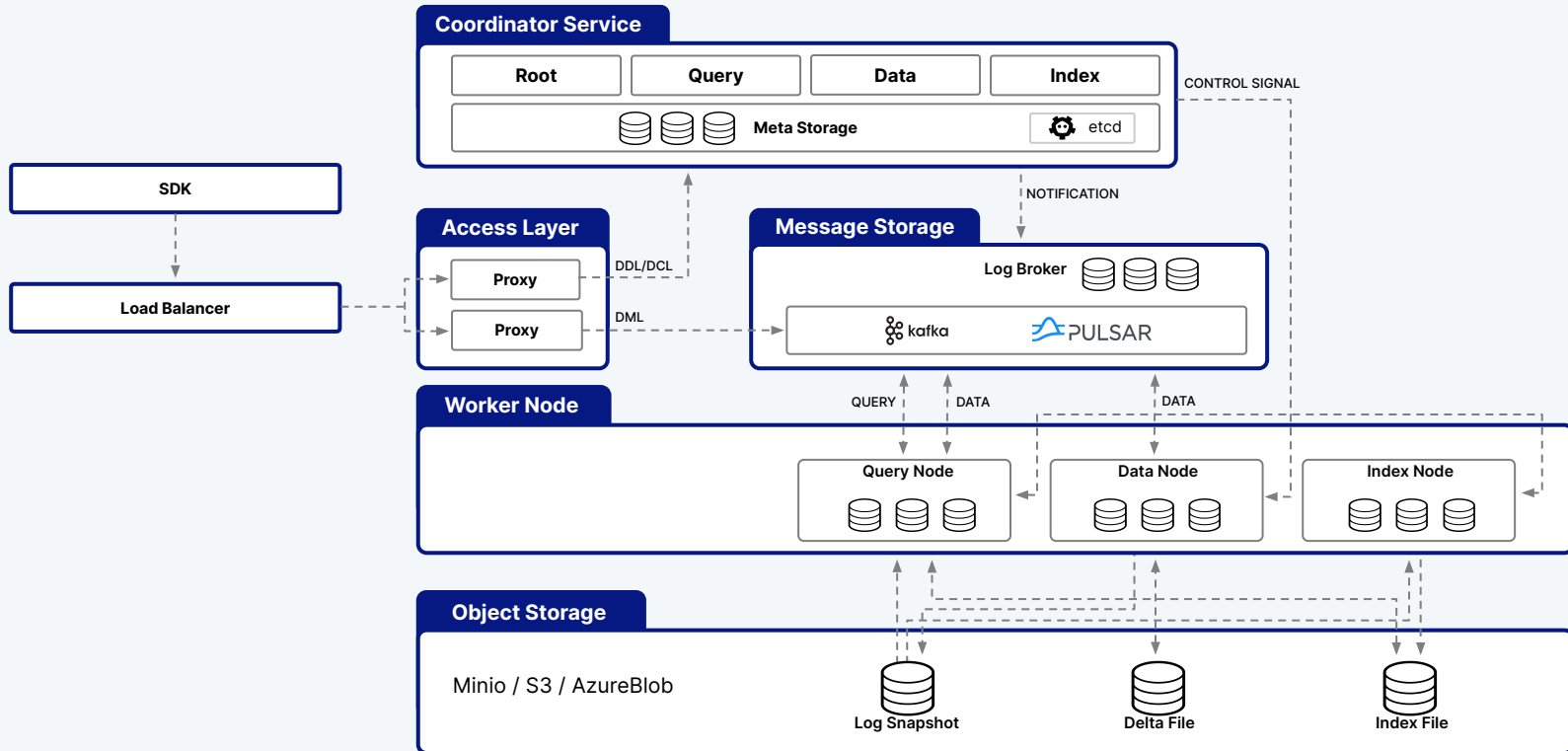
06 Demo

Give Milvus a Star!

zilliz

RAG AI Agent Architecture

Response

Tools

milvus

Request

Your Data

zilliz

Start building
with Zilliz Cloud today!
zilliz.com/cloud

zilliz