Effective RAG: Generate and Evaluate High-Quality Content for Your LLMs

Atindriyo Sanyal | Galileo
Yujian Tang | Zilliz

Speaker

Yujian Tang

Developer Advocate, Zilliz

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

 |

zilliz

Speaker

Atindriyo Sanyal

Co-Founder, Galileo

atin@rungalileo.io
https://www.linkedin.com/in/atinsanyal/

 |

zilliz

Outline:

1. **Introduction**
    a. Here are these use cases and how you can use LLMs to solve that problem
    b. Challenge of hallucination
2. **Intro RAG**
    a. That's where RAG comes in
3. **Challenges w/ RAG**
    a. Lack of metrics that measure if you have the right content and context
    b. Need a way to systematically measure if you have the right context
    c. Galileo offers consistent evaluation metrics across prompting and production monitoring

zilliz

Demo Flow: Go over a way to chat w/ Towards Data Science (here's how we'll retrieve data,

Goal: Chat with data taken from Towards Data Science

1. **Connect your workbook to Zilliz**
   a. Yujian has prepped workbooks w/ reasonable responses
   b. Uploaded on Zilliz cloud (will share file via Google)
2. **Conduct a run**
   a. In the workbook, show a user asking a question of the LLM
   b. **Question 1**: Explain Vector Embeddings to Me
      i. Answer 1 (in notebook): LLM responds with an answer derived from Towards Data Science
   c. **Question 2**: What is the best phone to buy
      i. Answer 2 (in notebook): The best phone to buy is the iPhone 15 Pro (this is clearly out of context as it's not derived from Towards Data Science)
3. **Pivot into Galileo (show Galileo UI)**
   a. For Question 2, Answer 2 is based on data from o outside Towards Data Science (not good)
      i. Show Groundedness score and hover on explanation
   b. Edit prompt and ask Question 2 again (can we edit the prompt in Galileo UI? If not, show in notebook)
      i. Expected output: "I don't have an output for this question"
      ii. Answer 3: Sorry don't have enough context, Galileo groundedness = 0

zilliz

CONTENTS

 |

zilliz

01

Why Use RAG?

zilliz

A Hallucination Problem

FORBES > BUSINESS

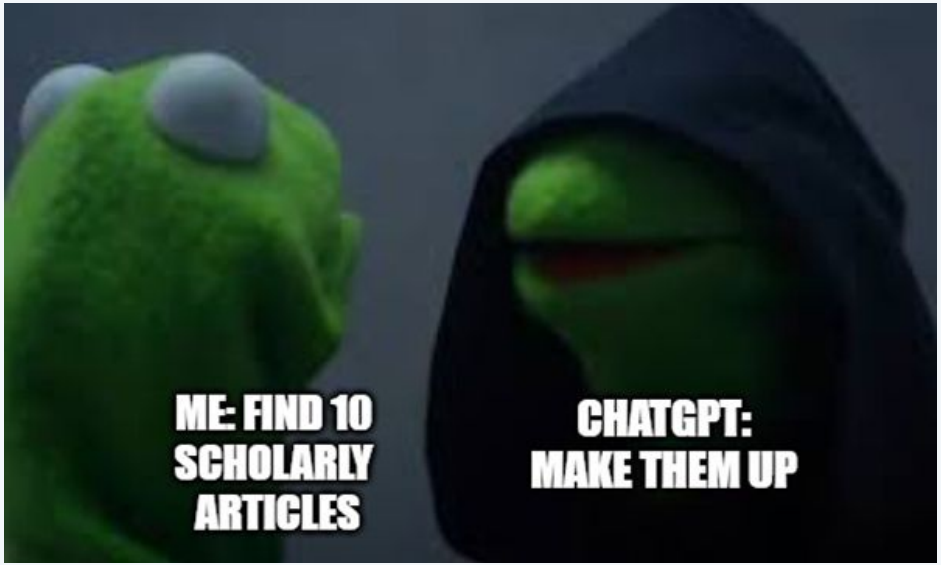
BREAKING

Lawyer Used ChatGPT In Court—And Cited Fake Cases. A Judge Is Considering Sanctions

Molly Bohannon Forbes Staff

I cover breaking news.

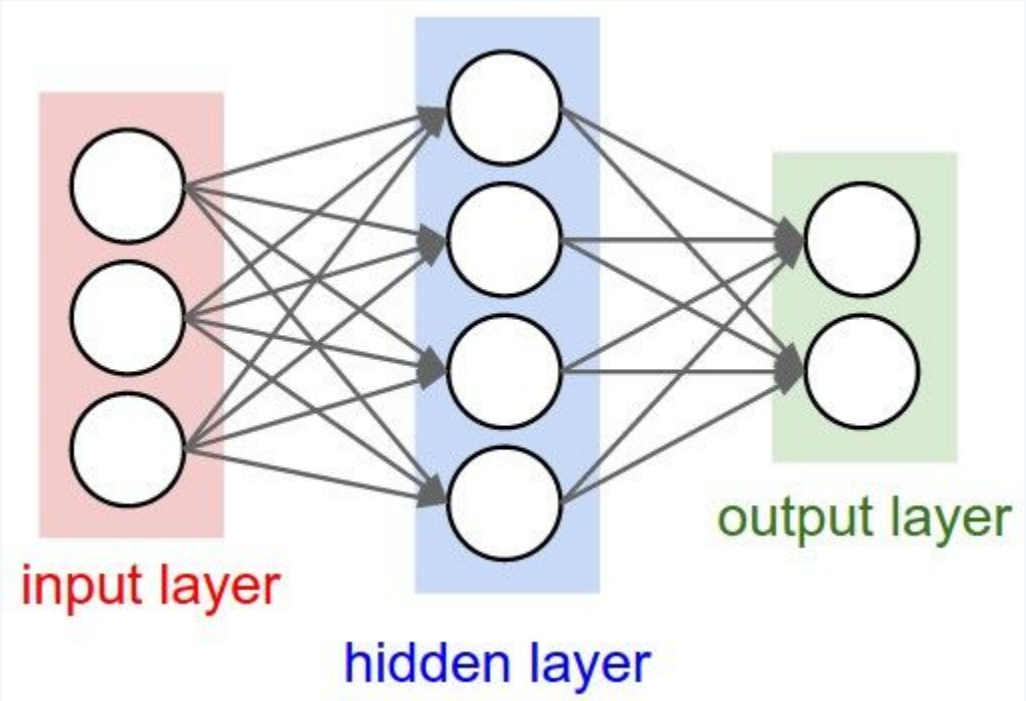
Follow



 |

zilliz

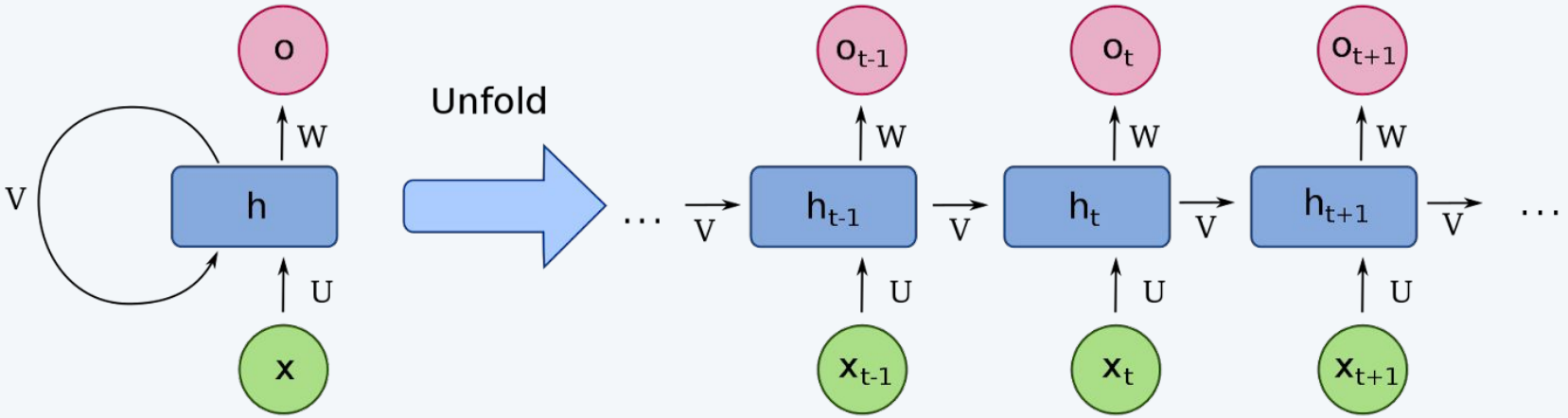
A Basic Neural Net



 |

zilliz

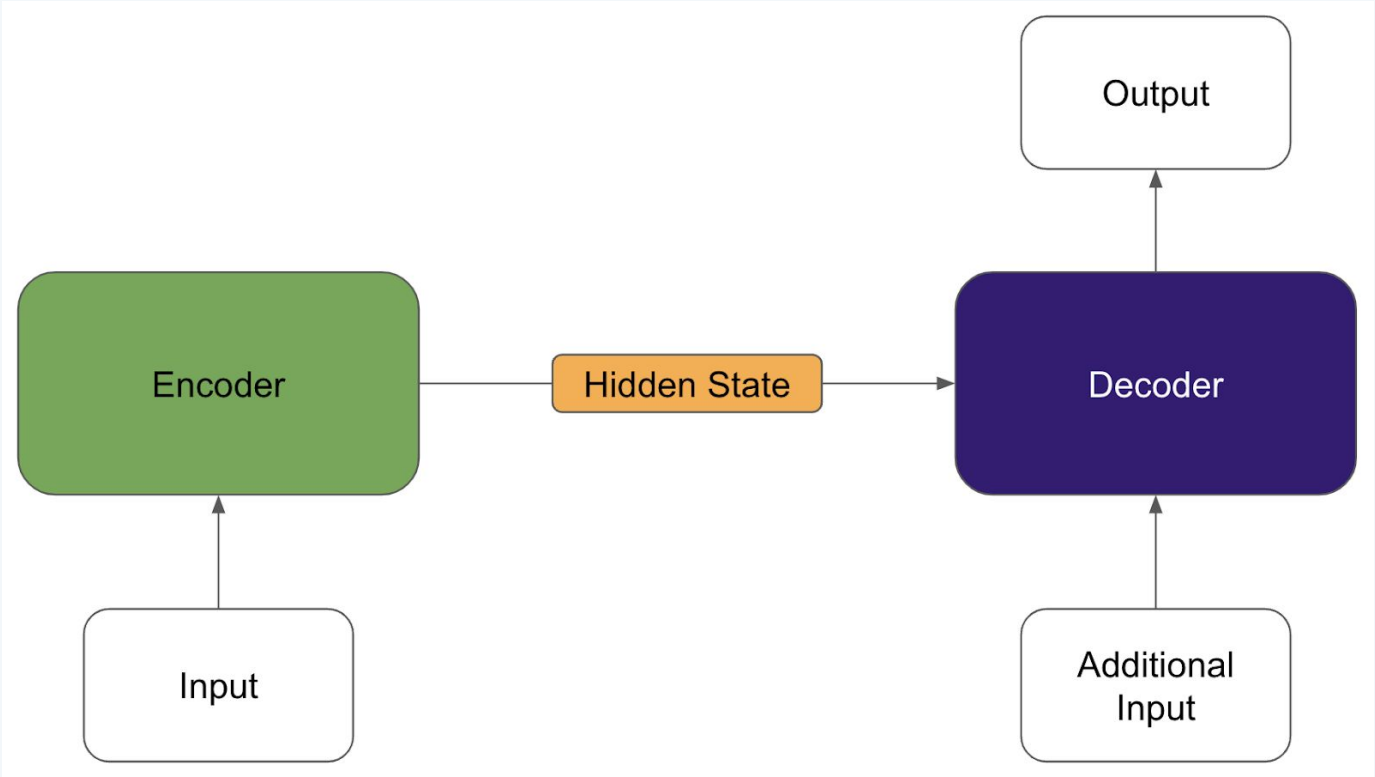
A Recurrent Neural Network



 |

zilliz

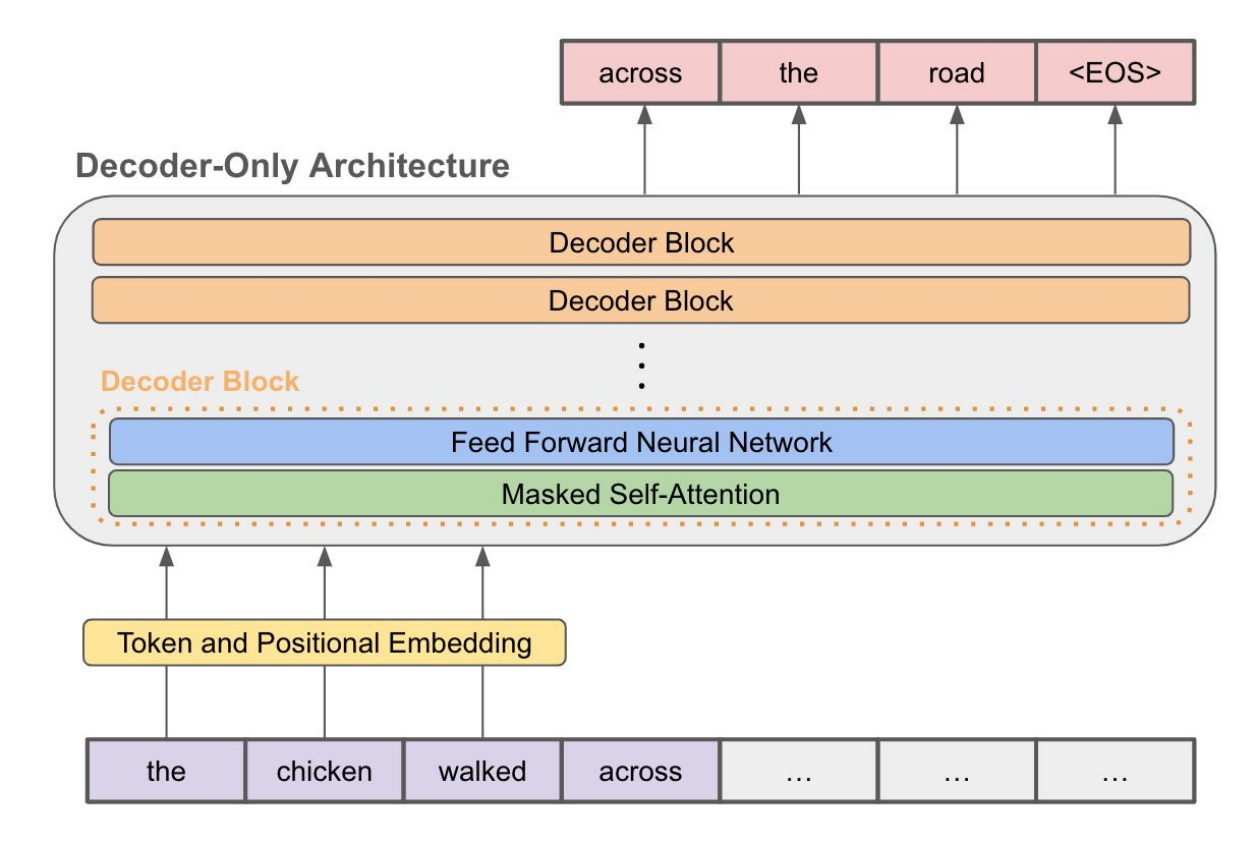
A Transformer Architecture



 |

zilliz

GPT Architecture



 |

zilliz

Takeaway

The reason ChatGPT hallucinates is because ...

It’s set up to predict a series of words (tokens)

 |

zilliz

02

How Can You Build Your RAG App?

zilliz

What is RAG?

Inject YOUR custom data on top of an LLM

Use similarity search to find the right data

 |

zilliz

Basic RAG Architecture

 |

zilliz

What’s a RAG tech stack look like?

CVP Stack

C: ChatGPT (or any other LLM)

- This can also be interpreted as the “processor” block for CVP

V: Vector database (e.g. Milvus)

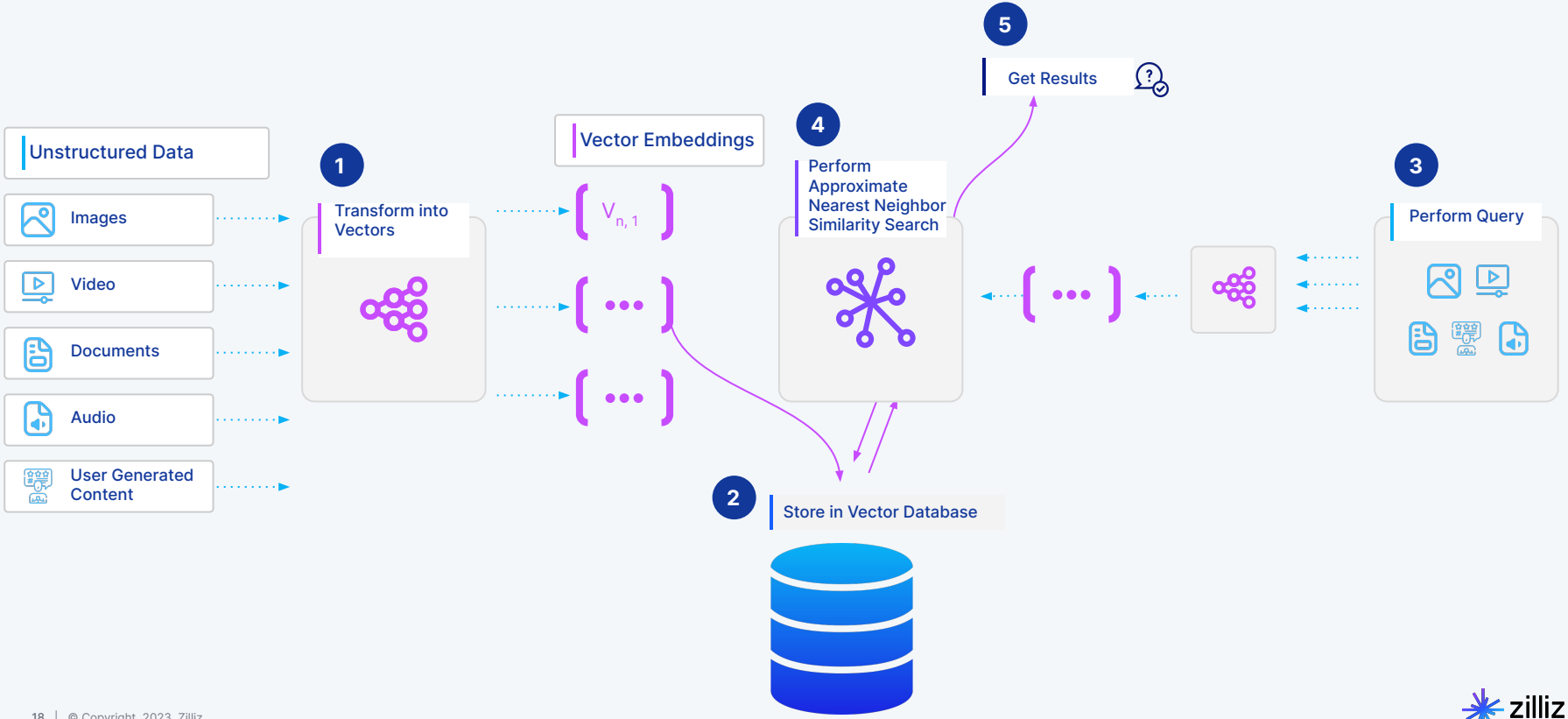
- Can also be interpreted as the “storage” block for CVP

P: Prompt-as-code (e.g. Haystack)

- Interface between processor and storage blocks

zilliz

How Similarity Search Works



 |

zilliz

Takeaway

A RAG App can be built like a computer

using an LLM for compute (CPU/GPU), a vector database for storage (hard drive), and prompt as code (interface)

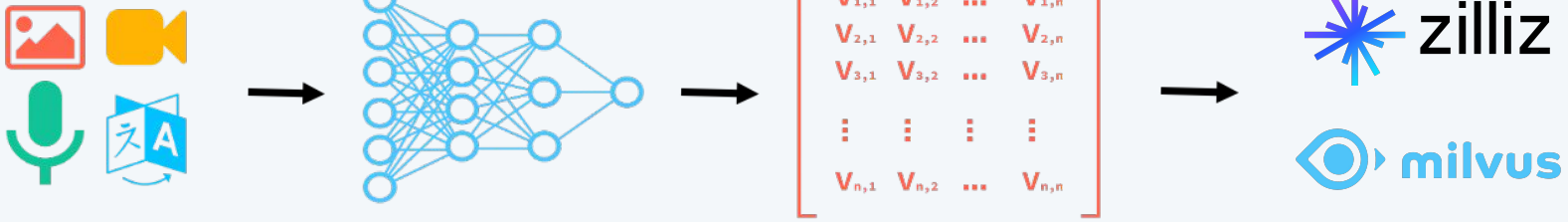
 |

zilliz

03

The Role of Vector Embeddings

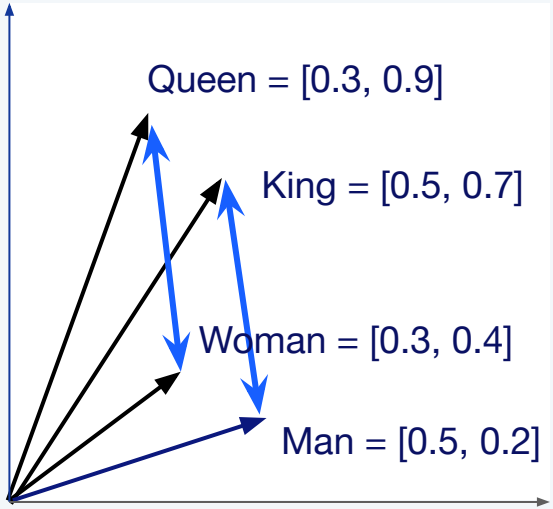
Where do Vectors Come From?

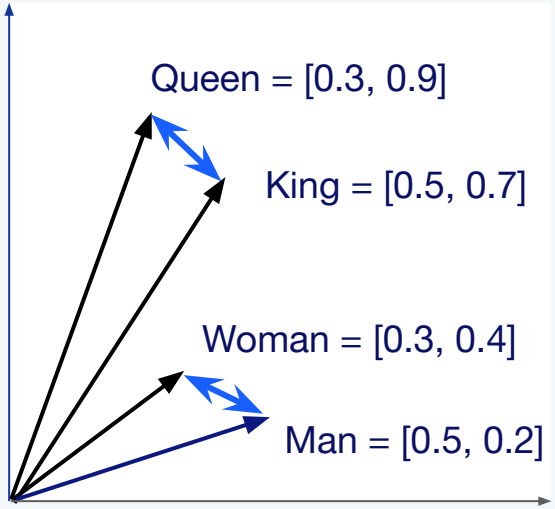


 |

zilliz

Semantic Similarity





Queen - Woman + Man = King

Queen = [0.3, 0.9]
\- Woman = [0.3, 0.4]

[0.0, 0.5]
\+ Man = [0.5, 0.2]

King = [0.5, 0.7]

Image from Sutor et al

 |

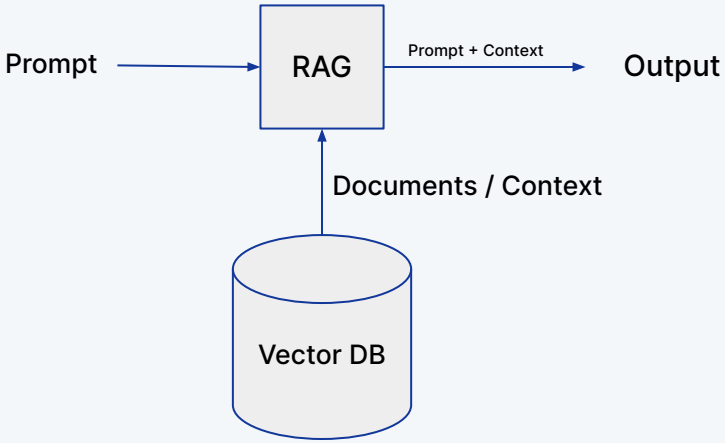
zilliz

04

Evaluating Your RAG Outputs Using Embeddings

zilliz

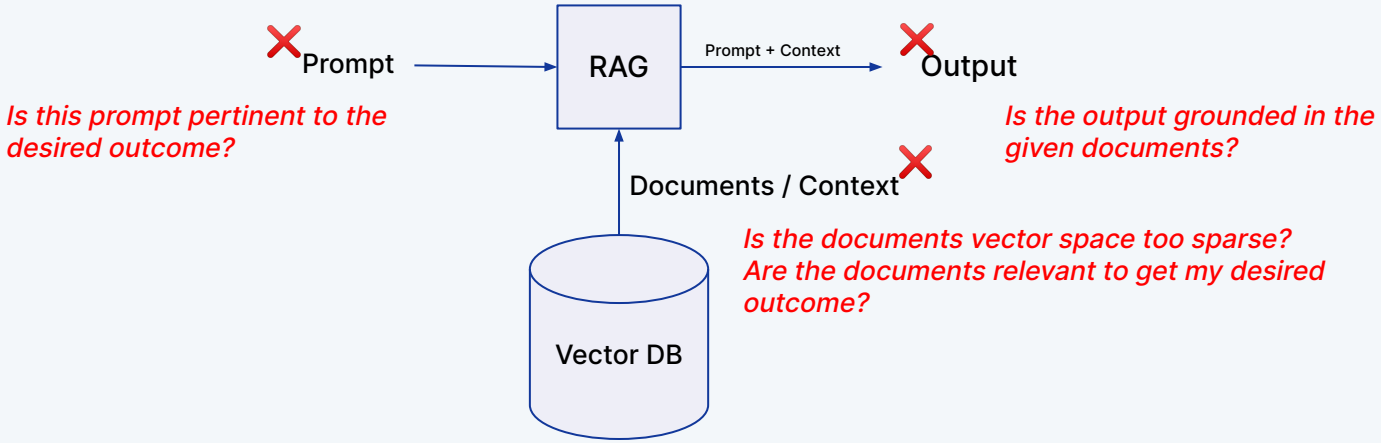
Evaluating Your RAG Outputs Using Embeddings



 |

zilliz

Evaluating Your RAG Outputs Using Embeddings

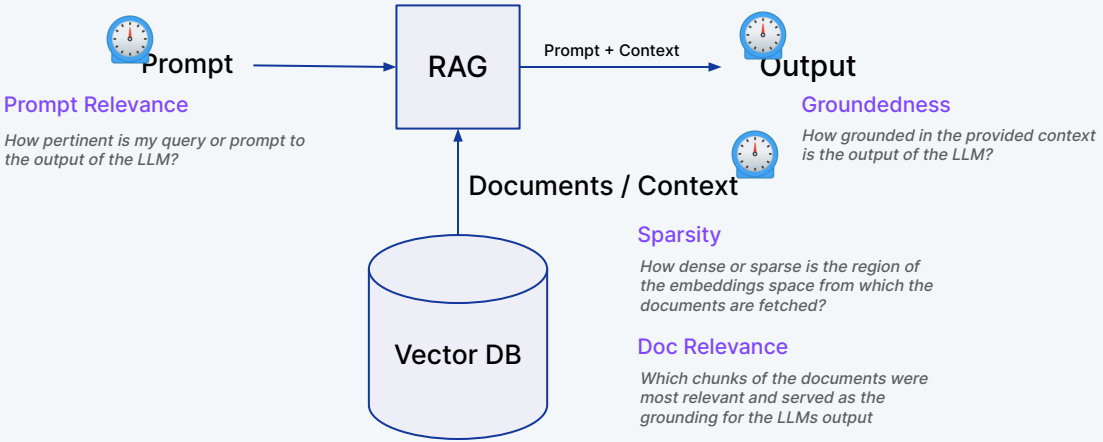


A 10K foot view of a RAG workflow

 |

zilliz

Evaluating Your RAG Outputs Using Embeddings

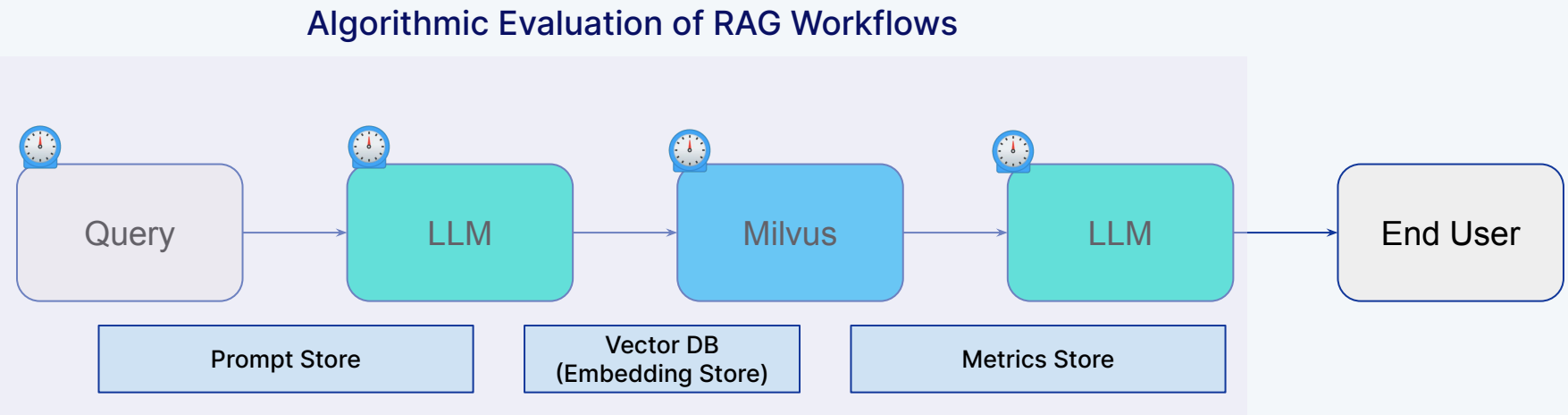


A 10K foot view of a RAG workflow

 |

zilliz

Evaluating Your RAG Outputs Using Embeddings



 |

zilliz

05

Demo

zilliz

THANK YOU

zilliz

Galileo

05

Appendix