

# AI AGENTS CHALLENGE

20.05.2025

# AGENDA



# AGENDA

10:00 – 10:30



Introduction Reply and AI Software Engineering

10:30 – 12:30



Hands-On introduction in RAG

12:30 – 13:00



Reply AI Coding Challenge

13:00 – 14:00



*Interactive Lunch Break*

14:00 – 16:00



Challenge Time

16:00 – 17:00



Solution presentation and winner announcement





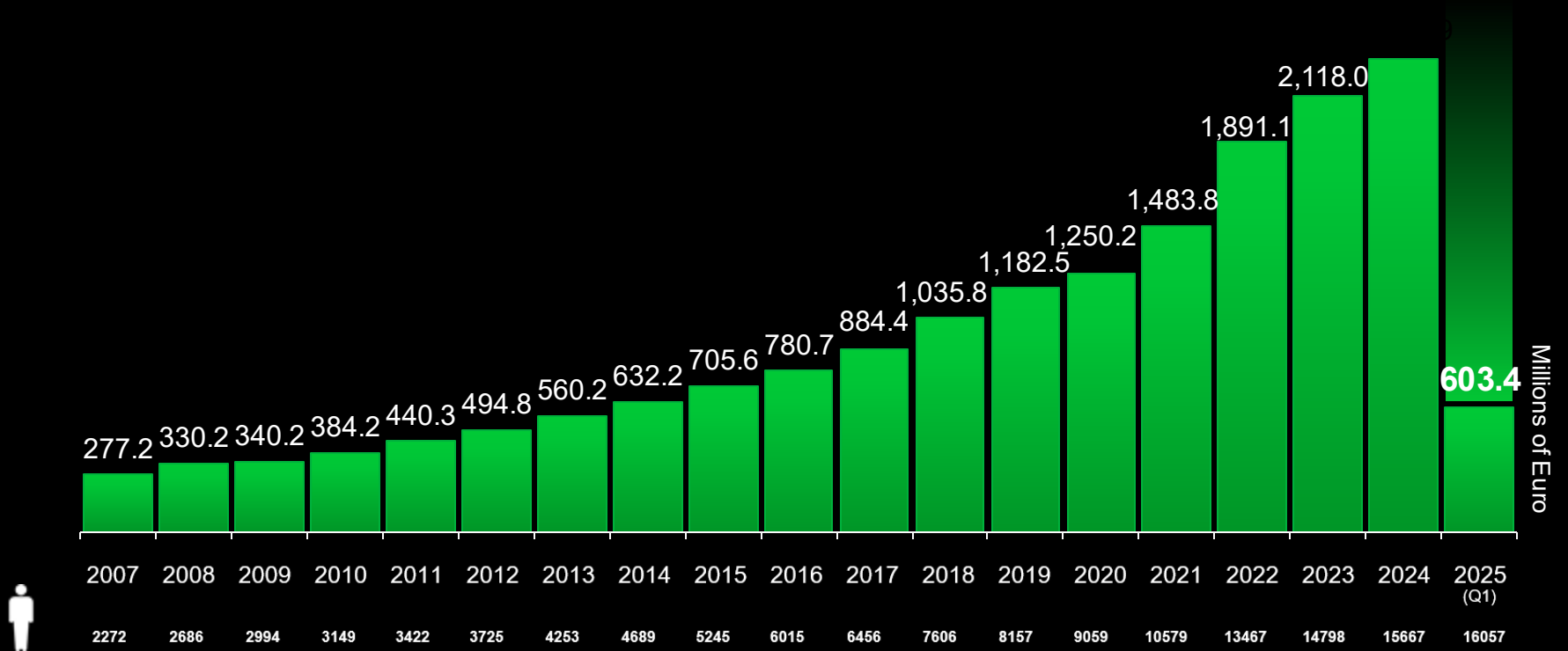
**REPLY**





**To excel in helping our customers exploit relevant innovation brought about by economic changes and driven by internet technologies.**

# REVENUE & PEOPLE



**US**  
Atlanta, Chicago, Detroit, Jacksonville,  
Kansas City, Philadelphia,  
New York, Seattle

**BRAZIL**  
Belo Horizonte, São Paulo

**UK**  
Manchester, **LONDON**, Sheffield,  
Edinburgh, Bristol

**FRANCE**  
Lille, Paris

**BENELUX**  
Amsterdam, Antwerp, Brussels,  
Luxembourg, The Hague

**SWITZERLAND**  
Zurich

**MOROCCO**  
Casablanca

**GERMANY**  
Berlin, Bremen, Dortmund, Düsseldorf, Frankfurt, **GÜTERSLOH**,  
Hamburg, Karlsruhe, Köln, Lübeck, Minden, München, Potsdam,  
Regensburg, Stuttgart

**POLAND**  
Katowice

**CROATIA**  
Zagreb

**ROMANIA**  
Bucharest

**AUSTRIA**  
Vienna, Innsbruck

**ITALY**  
Bari, Bologna, Firenze, Genova,  
Milano, Padova, Parma, Roma,  
**TORINO**, Treviso, Trieste, Verona

**INDIA**  
Kochi

**CHINA**  
Beijing, Nanjing

**SINGAPORE**  
Singapore



# REPLY SERVICES

## PEOPLE

DIGITAL HUMANS

COPILOTS

WORKPLACES

## MACHINES

ROBOTICS

AUTONOMOUS  
THINGS

IOT & CONNECTED  
PRODUCTS

## DIGITAL EXPERIENCE

PRODUCT DESIGN  
BRAND EXPERIENCE

DIGITAL MARKETING

IMMERSIVE  
EXPERIENCES

## ENTERPRISE PLATFORMS

ADVANCED  
ANALYTICS

INDUSTRY SPECIFIC  
PLATFORMS

APPLICATION SUITES  
& CX PLATFORMS

## AI

## FOUNDATIONS

COMPUTING PLATFORMS

DATA PLATFORMS

WEB 3.0

CYBERSECURITY

NETWORK

3D & SPATIAL COMPUTING





The image displays a collection of 100 corporate logos, organized in a 10x10 grid. The logos represent a wide variety of industries and global brands, including technology, finance, retail, and manufacturing. Some of the recognizable logos include Amazon, Google, Microsoft, Apple, and many others. The logos are presented in their original colors and designs, set against a plain white background.



# PARTNERS ECOSYSTEM



Platinum Partner



Global Partner



Cloud Premier  
Partner



Global Partner



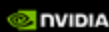
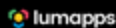
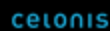
Global Partner



EMEA Strategic  
Partner



Gold  
Partner



# MAIN AWARDS & ACHIEVEMENTS

2025



GENERATIVE AI  
SPECIALIZATION IN THE  
GOOGLE CLOUD PARTNER  
ADVANTAGE PROGRAM  
Go Reply,  
Machine Learning Reply

2025



SALESFORCE  
IMPLEMENTATION PARTNER  
OF THE YEAR  
Arlanis Reply

2025



FINANCIAL TIMES' LEADING  
MANAGEMENT CONSULTANTS  
ANNUAL UK LIST  
Reply

2025



PARTNER OF THE YEAR  
FOR COUNTRY – UK/IE  
Go Reply

2025



MULTI-YEAR STRATEGIC  
COLLABORATION  
AGREEMENT TO ACCELERATE  
INNOVATION IN GENAI  
Reply

2025



GARTNER® MARKET  
GUIDE FOR  
STRATEGIC WEBSITE  
AGENCIES  
Sagepath Reply

2024



"BEST DIGITAL PARTNER"  
AWARD NURSETECH  
INNOVATION CHALLENGE  
Solidsoft Reply

2024



VISIONARY IN MAGIC  
QUADRANT FOR WMS  
Reply

2024



LEADER "SALESFORCE MULTICLOUD  
IMPLEMENTATION AND INTEGRATION SERVICES  
FOR LARGE ENTERPRISES" GERMANY  
LEADER "SALESFORCE IMPLEMENTATION  
SERVICES FOR MARKETING AUTOMATION  
FRANCE & GERMANY"  
Reply

2024



WINNER  
DIGITAL FACTORY  
AWARD  
Axulus Reply

2024



ADVANCED INDUSTRIAL  
ROBOTIC APPLICATIONS  
(AIRA) CHALLENGE  
WINNER  
Roboverse Reply

2024



BVDW INTERNET  
AGENCY RANKING 2nd  
PLACE  
Reply

2024



LOGIMAT BEST  
PRODUCT AWARD  
Lea Reply

2024



NONPROFIT  
MICROSOFT PARTNER  
OF THE YEAR  
Valorem Reply

2024



EUROPEAN LAND ROBOT  
TRIAL 2024 "BEST  
PERFORMANCE" AWARD  
Roboverse Reply

2024



ORACLE CLOUD MANAGED  
SERVICE PROVIDER  
Red Reply, Technology Reply

2024



MICROSOFT PARTNER OF  
THE YEAR FOR ITALY  
Cluster Reply

2024



NC DIGITAL AWARDS  
Triplesense Reply,  
Bitmama Reply

2024



ASANA AMER PARTNER  
OF THE YEAR  
Spur Reply

2024



LEADER IN DIGITAL  
EXPERIENCE SERVICES  
Reply

2024



ORACLE PARTNER AWARDS,  
CLOUD/TECHNOLOGY EUROPE  
SOUTH INNOVATION  
Technology Reply

2024



ADOBE EXPERIENCE  
MANAGER SITES  
SPECIALIZATION  
Sagepath Reply

2024



"BEST CROSS-BORDER ROLL OUT"  
ECOMMERCE AWARDS  
Portaltech Reply

2024



FRAUNHOFER VALIDATION  
FOR MATERIAL FLOW AND  
LOGISTICS (IML)  
LEA Reply

2024



SAP QUALITY AWARDS FOR  
CUSTOMER SUCCESS  
Syskoplan Reply

2024



RECOGNITION IN  
FORRESTER'S SAP  
SERVICES LANDSCAPE

2024



PARTNER OF THE  
YEAR FOR DEVOPS  
SPECIALISATION  
Go Reply

2024



AWS GENERATIVE AI  
COMPETENCY  
Reply





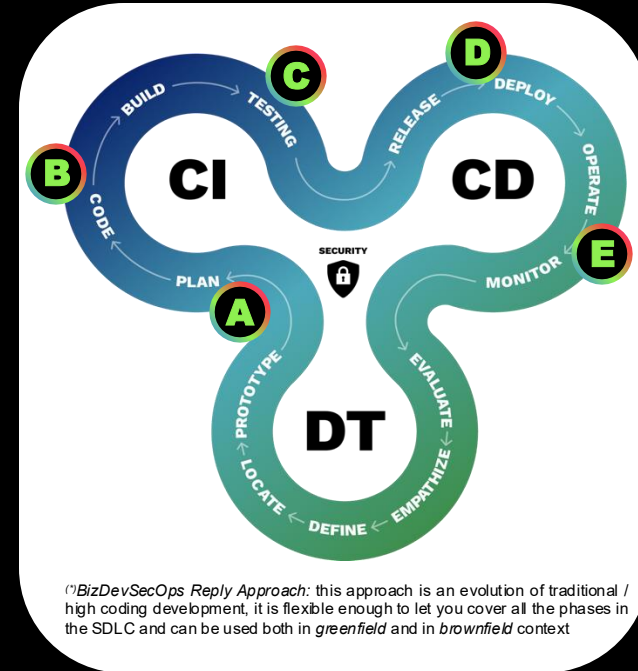
# AI SOFTWARE ENGINEERING



# BOOST SDLC WITH GEN AI

The use of generative AI applied either in a greenfield project or in a Legacy ecosystem can provide support for:

1. **Extracting valuable insights** from different sources (knowledge as asset)
2. **Improve operational efficiency in the software life cycle (\*)**:
  - A Analysis & Design**: requirement analysis, user story generation, Sequence, Flow Diagram Generation, Data Model Authoring, UX Design Assistant
  - B Coding**: code generation, explain, debugging, translation, improvement on consistency
  - C Code Quality**: test cases writing, generation and execution.
  - D Release&Deployment**: CI/CD deployment generation, automation scripting writing assistant
  - E Operate&Monitoring**: performance monitoring, tickets auto routing and solution patterns finding
3. We are also **experimenting support** for **legacy modernization**



# X STORY ASSISTANT



## X Story Assistant

By Daniel Hummel

Write User Stories for the X Team.

How can this user story be improved?

Please rewrite and improve this User

Send message



The user of the CRS like to receive fixes for the sonarqube findings.





# PLANNING AND EXECUTION

## TIME CONSUMPTION

WRITE A  
USER  
STORY

WRITE  
ACCEPTANC  
E CRITERIA

VERIFY  
WITH THE  
DOR

ESTIMATE  
EFFORT

DEFINE  
TEST  
CASES

AND MORE..



# CODE GENERATION SUPPORT



```
1 import math
2
3 # Compute the distance between two points on earth
4 def distanceEarth(lat1, lon1, lat2, lon2):
5     # Convert degrees to radians
6     lat1 = math.radians(lat1)
7     lat2 = math.radians(lat2)
8     lon1 = math.radians(lon1)
9     lon2 = math.radians(lon2)
10    # Compute using the Haversine formula
11    dlon = lon2 - lon1
12    dlat = lat2 - lat1
13    a = math.pow(math.sin(dlat/2), 2) + math.cos(lat1) * math.cos(lat2) * math.pow(math.sin(dlon/2), 2)
14    c = 2 * math.asin(math.sqrt(a))
15    # 6367 km is the radius of the Earth
16    km = 6367 * c
17    return km
18
19 # Compute the distance between two points on the moon
```





# LINT AGENT



st\_code\_quality/compare/main...source\_branch

ality

Search Type to search

Actions Projects Security Insights Settings

## Open a pull request

Create a new pull request by comparing changes across two branches. If you need to, you can also [compare across forks](#). [Learn more about diff comparisons here.](#)

base: main ← compare: source\_branch X Can't automatically merge. Don't worry, you can still create the pull request.

Source branch #14  
test

View pull request



Add a title

Source branch

Add a description

Write Preview

The pull request creates two new files that do not have good code quality (no exception handling, limited readability due to unnecessary lines of code).

Markdown is supported Paste, drop, or click to add files

Reviewers

No reviews

Assignees

No one—assign yourself

Labels

None yet

Projects

None yet

Milestone

No milestone

Development

Use [closing keywords](#) in the description to automatically close issues

Helpful resources

[GitHub Community Guidelines](#)

Remember, contributions to this repository should follow our [GitHub Community Guidelines](#).

6 commits

3 files changed

1 contributor

Commits on Mar 13, 2024

Inefficient code; luiz

Timokubera committed on Mar 13

9486c38 <>

more changes

Timokubera committed on Mar 13

0e9378d <>

# UNIT TEST AGENT



```
1 import math
2
3 # Compute the distance between two points on earth
4 def distanceEarth(lat1, lon1, lat2, lon2):
5     # Convert degrees to radians
6     lat1 = math.radians(lat1)
7     lat2 = math.radians(lat2)
8     lon1 = math.radians(lon1)
9     lon2 = math.radians(lon2)
10    # Compute using the Haversine formula
11    dlon = lon2 - lon1
12    dlat = lat2 - lat1
13    a = math.pow(math.sin(dlat/2), 2) + math.cos(lat1) * math.cos(lat2) * math.pow(math.sin(dlon/2), 2)
14    c = 2 * math.asin(math.sqrt(a))
15    # 6367 km is the radius of the Earth
16    km = 6367 * c
17    return km
18
19 # Compute the distance between two points on the moon
```



# MERGE CONFLICT RESOLVER



The screenshot displays a GitHub pull request titled "pull\_request\_merge\_conflict" for the repository "Timokubera/pull\_request\_merge\_conflict". The pull request is in a state of conflict, as indicated by the "pull\_request\_merge\_conflict" tab in the browser and the "pull\_request\_merge\_conflict / Calculator.java" file name. The file "Calculator.java" is shown with a conflict resolution interface. The code is as follows:

```
1 import java.util.List;
2
3 public class Calculator {
4     public int calculateSum(List<Integer> numbers) {
5         return numbers.stream().mapToInt(Integer::intValue).sum();
6     }
7 }
```

The interface includes a "Files" sidebar on the left with a search bar and a "Switch branches/tags" dropdown. The "Blame" tab is selected, showing the commit history for the file. The "Symbols" panel on the right shows the class "Calculator" and the method "calculateSum".

# 3<sup>RD</sup> LEVEL SUPPORT

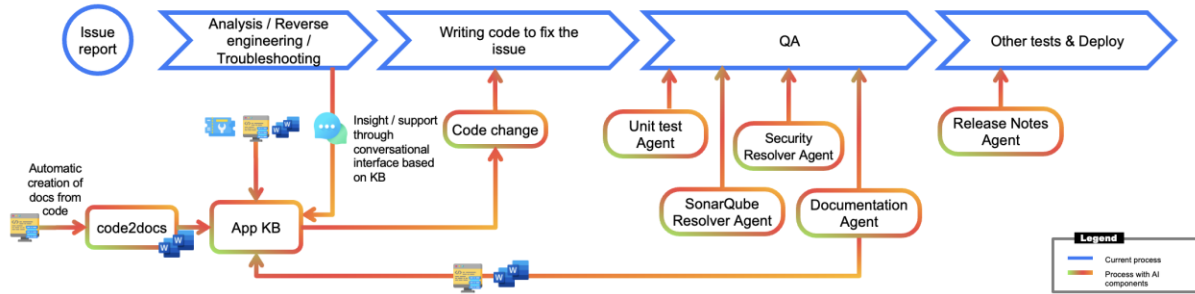
## AI-ENHANCED ERROR MANAGEMENT PROCESS

- Longer Root cause identification time
- True subject matter experts required
- Unsatisfied customers

### Challenge

- AI supported root cause analysis (RCA)
- AI supported fixing the found issue
- Deployment to AWS environment

### Solution



### Solution overview



Portability



Hallucination resistance



RCA support



Multiple integrations



CI/CD



OpenAI API integration



LangChain



AWS code pipeline



Kubernetes



Confluence API

### Key features



Faster root cause analysis



Consistent Quality and Documentation



AI driven development & testing



Operation efficiency

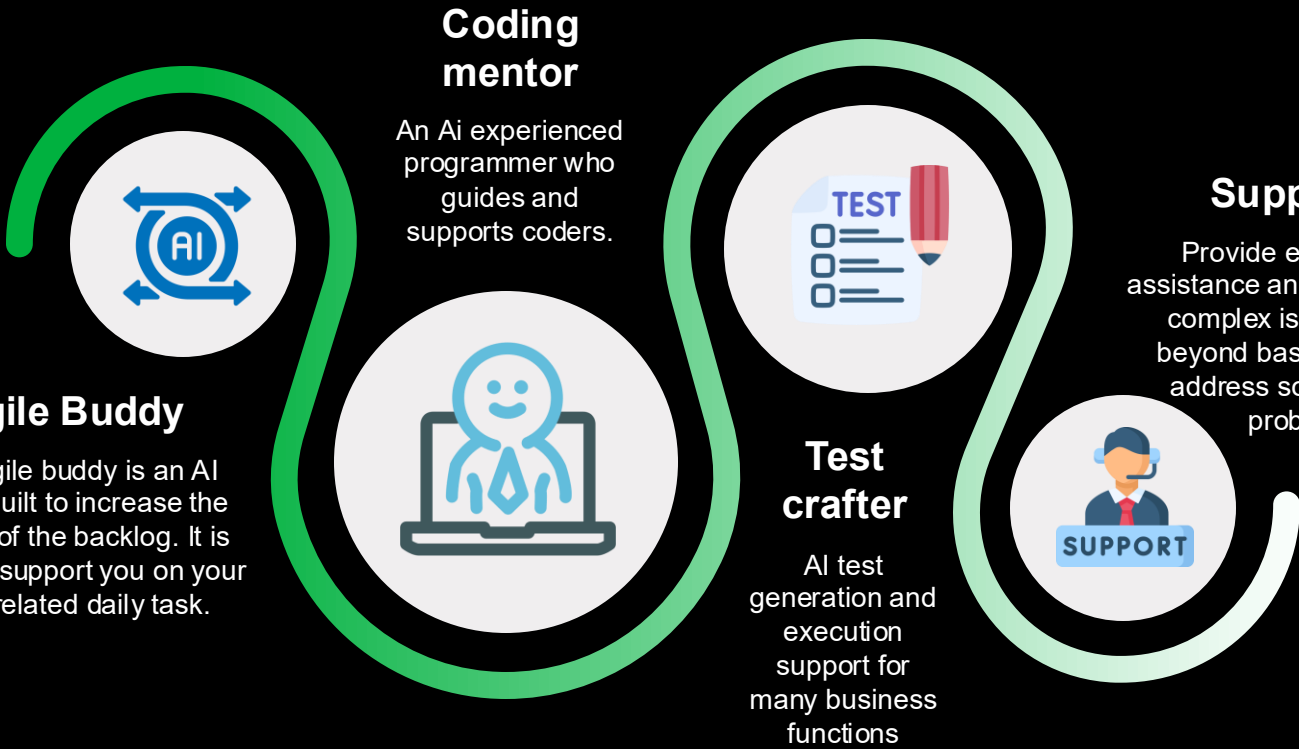
### Tools

### Benefits



# STARTING POINT

## AN ARTIFICIAL INTELLIGENCE TEAM OF AGENTS



# KICODE REPLY

**GENERATIVE AI ACCELERATOR FOR THE SOFTWARE DEVELOPMENT LIFE CYCLE BY KI REPLY**

Ki Reply offers a task-driven autonomous agent system powered by Generative AI, enabling rapid, fully automated software development.

It understands natural language commands, breaks them into subtasks, and delegates them to various AI agents, streamlining the entire Software Development Life Cycle (SDLC) from idea to deployment in minutes.

A registration form titled "Register With Us" with fields for Username, Email, Password, and Confirm Password, and a Submit button.

Register With Us

Username  
Enter username

Email  
Enter email

Password  
Enter password

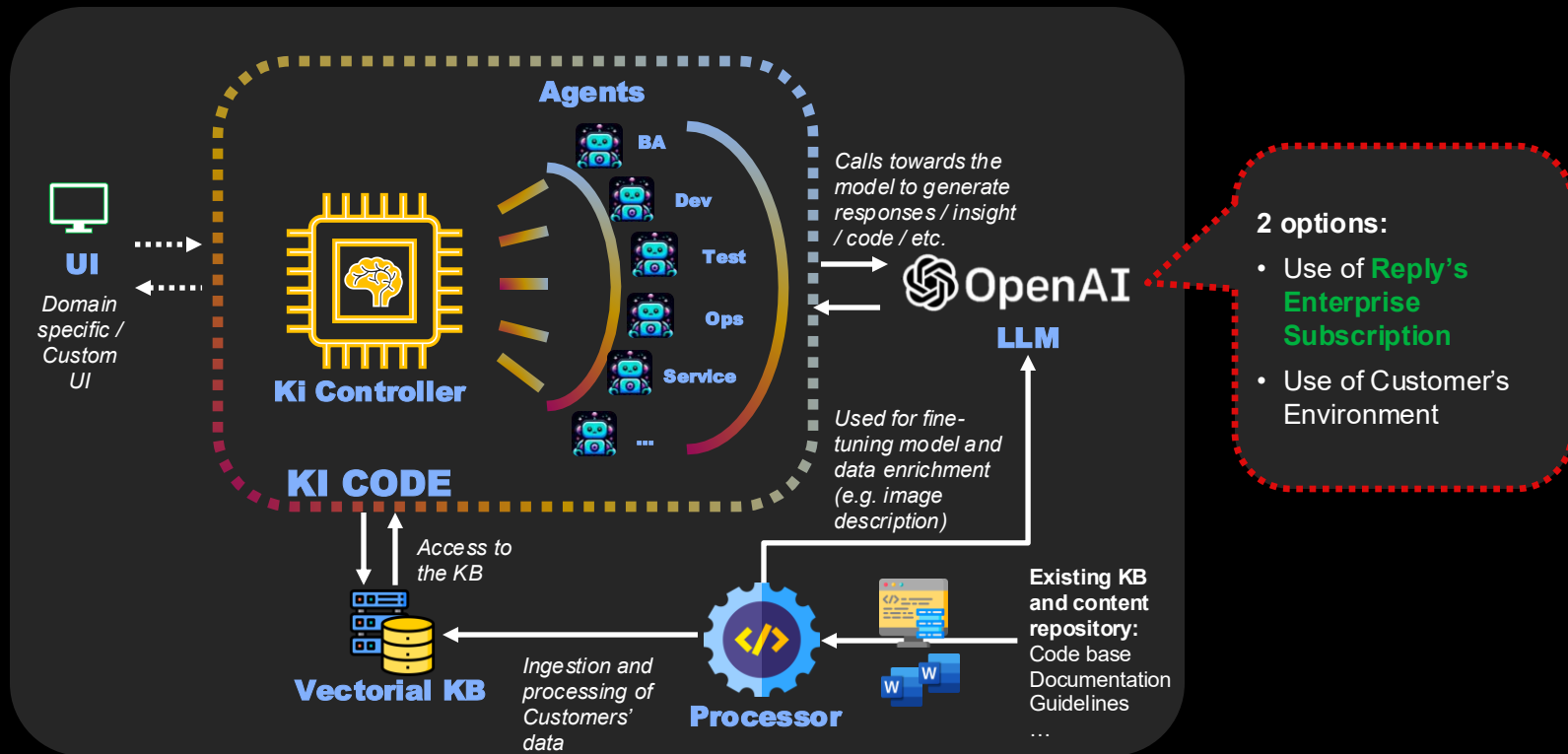
Confirm Password  
Enter password again

Submit

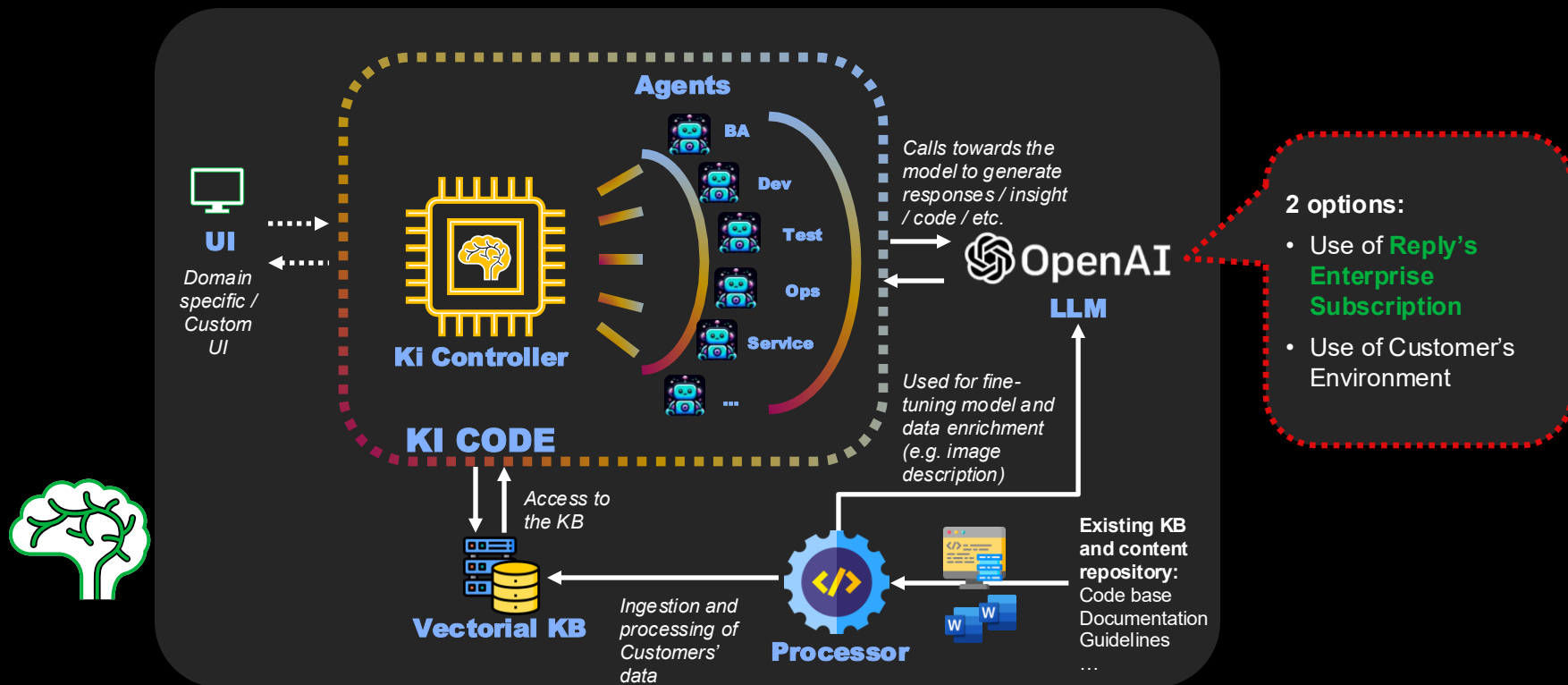
<https://www.reply.com/de/artificial-intelligence/kicode-reply>



# KICODE: MULTI-AGENT APPROACH



# KICODE: MULTI-AGENT APPROACH



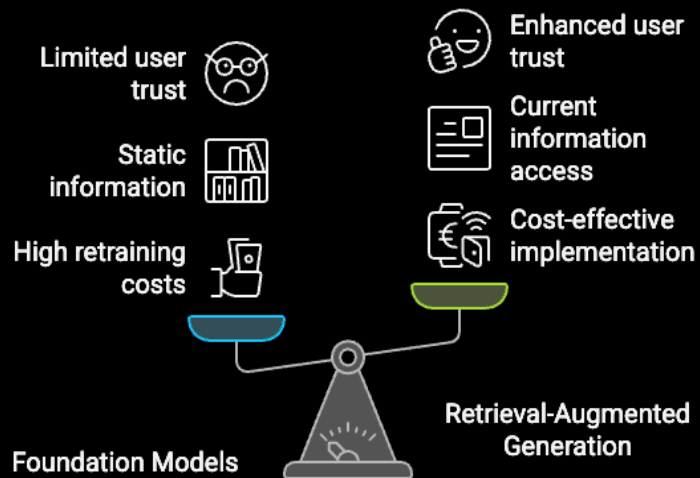




# **THEORETICAL ONBOARDING**



# RAG, WHY?

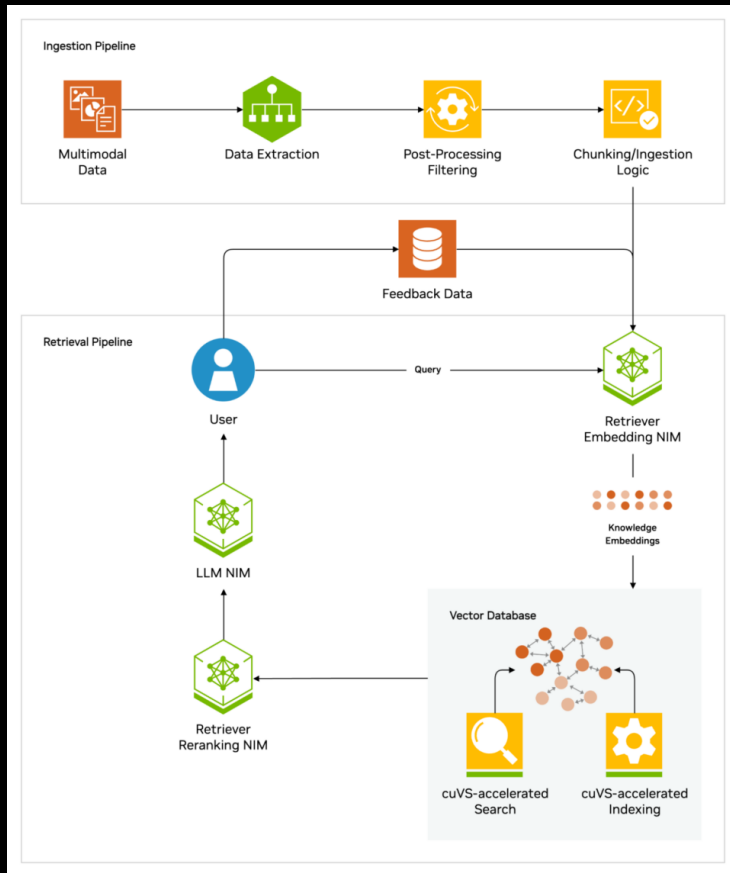


**RAG offers cost-effective, current, and trustworthy AI solutions.**

Made with  Napkin



# RETRIEVAL-AUGMENTED GENERATION (RAG)



"

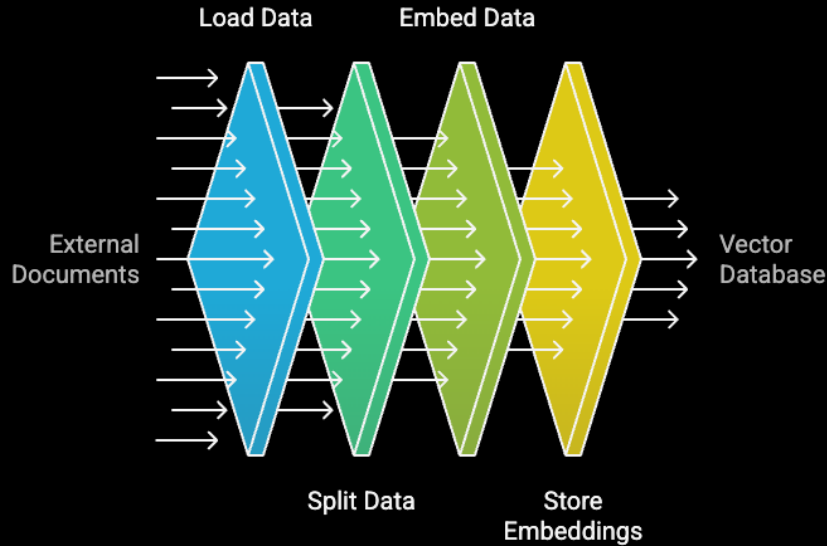
A technique for enhancing the accuracy and reliability of generative AI models with information fetched from specific and relevant data sources

"



# RAG: INGESTION

## Data Ingestion Process



"

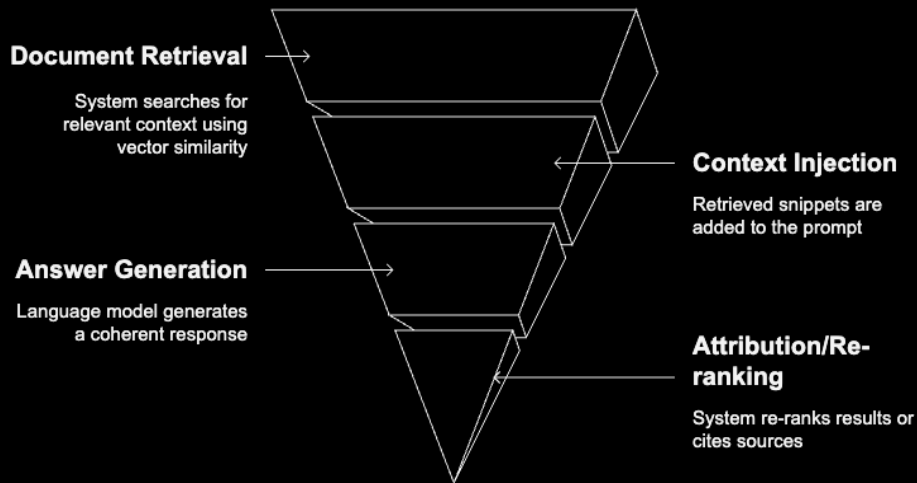
A robust RAG system relies on a well-designed ingestion pipeline, as it retrieves answers based on external context. The better the ingestion quality, the more accurate the output.

"



# RAG: CONTEXT TUNING

## Enhanced Query Processing Funnel



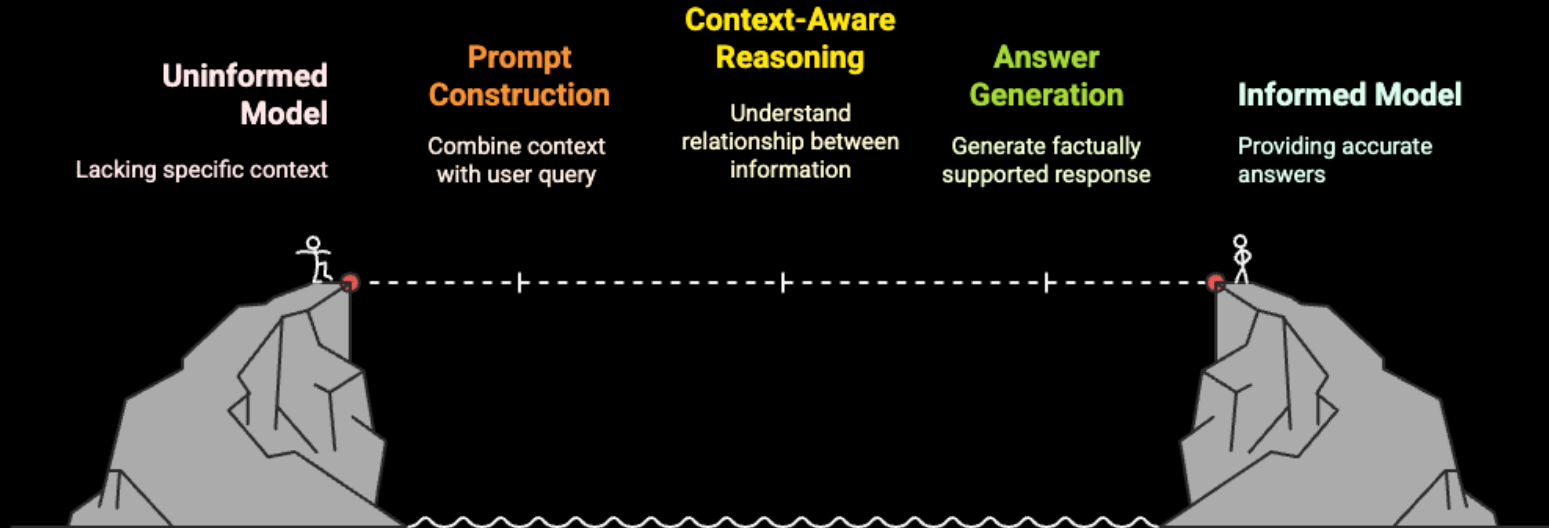
"  
Providing Context for a better  
LLM response  
"



# RAG

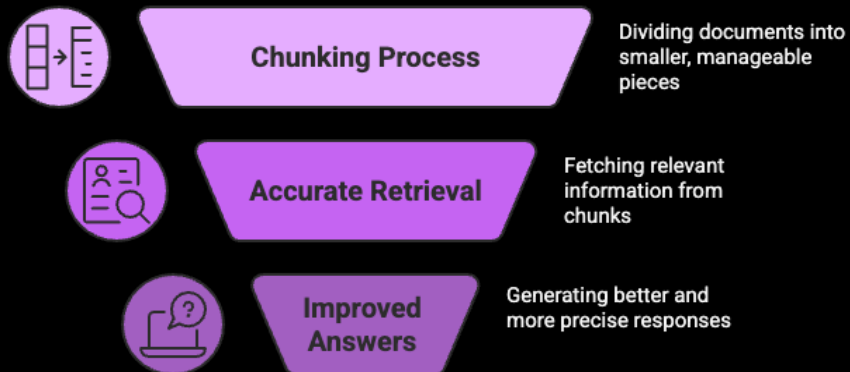
## GENERATION

Generating Context-Aware Answers



# CHUNKING

## Enhancing Information Retrieval through Chunking



Chunking refers to the process of splitting documents into smaller, manageable pieces (chunks) before storing them in a vector database for retrieval



# SEMANTIC SEARCH

Example:

- 1 **User query:** "How can I pay taxes on freelance income?"
- 2 **Semantic match:** "Freelancers must file estimated taxes quarterly in the U.S."
- 3 **Keyword match (bad):** Would match "income", "taxes" — but might miss relevant context like "freelance" or "quarterly payments".

"

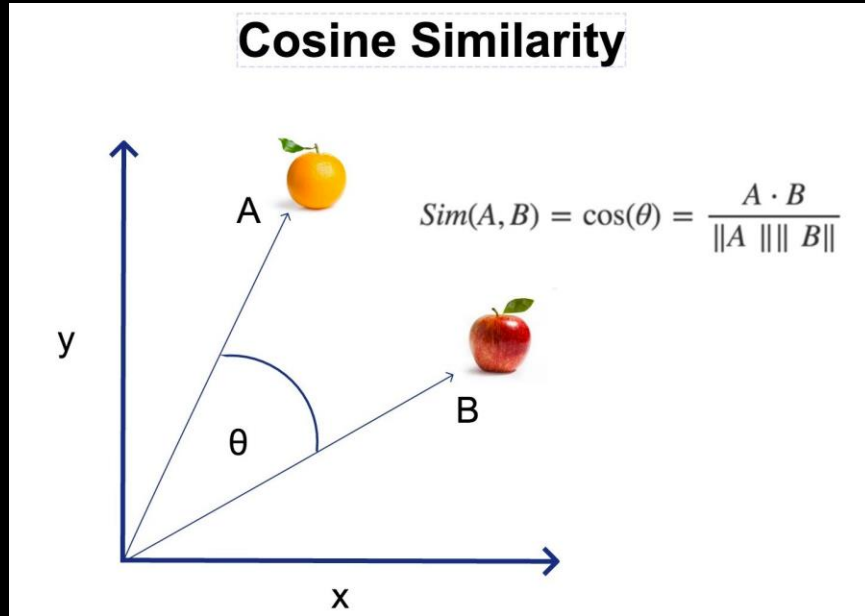
**Semantic search** is retrieving information based on the **meaning** of the query rather than matching keywords.

"





# COSINE SIMILARITY



- **Query:** "How to scrape data from a website?"
- **Chunk A:** "This code uses BeautifulSoup to scrape HTML content from web pages."
- **Chunk B:** "The database stores product reviews in JSON format."
- $\text{cosine\_similarity}(Q, A) \rightarrow$  **High score (close to 1)**
- $\text{cosine\_similarity}(Q, B) \rightarrow$  **Low score (close to 0)**



# EMBEDDINGS

## Advantages of RAG in Different Contexts

1

### Efficient Retrieval

Supports fast top-k nearest neighbor search with limited context.



2

### Semantic Matching

Understands meaning beyond keywords for efficient retrieval.



3

### Domain Adaptability

Adapts to different models with low contextual understanding.



4

### LLM Grounding

Improves factuality by providing relevant context but lacks efficiency.



"

An **embedding** is a numerical vector that represents the **meaning** of a piece of text in a high-dimensional space.

"



# VECTORS EXAMPLE

Text chunk: "Scrape data from a website using Python"

Embedding vector:  $[0.3, 0.7, 0.1, 0.9, \dots]$  (high dimensional)

Query: "How can I extract web data with code?"

Query vector:  $[0.32, 0.72, 0.12, 0.88, \dots]$

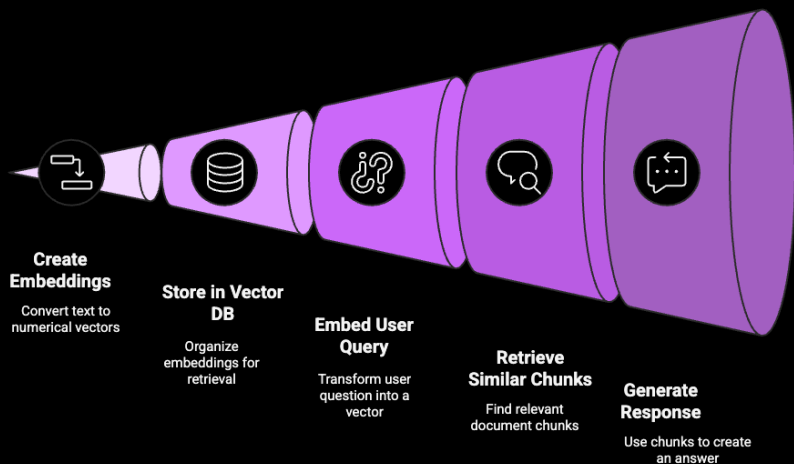
Similarity measure: Cosine similarity between vectors

Usage: Retrieve top-matching chunks for grounding LLM



# VECTOR DATABASES

RAG Process Funnel



Made with Napkin

"

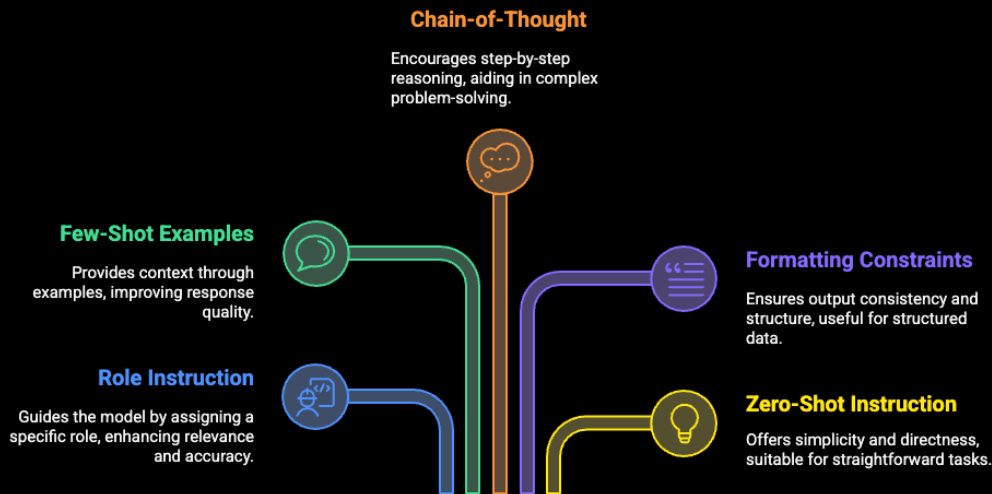
A **vector database** stores and indexes **embeddings** — high-dimensional numerical representations of text, images, or other data — and allows for **fast similarity search**.

"



# PROMPT ENGINEERING

Which prompt engineering technique should be used?



Poor prompt:

*"Summarize this."*

Better prompt:

*"Summarize the following academic article in 3 bullet points highlighting its key findings and conclusion."*



# LANGCHAIN

## Agents

Enables LLMs to perform actions like web searches and file writing

## Code Assistants

Aids developers in coding tasks with intelligent suggestions

## RAG Pipelines

Integrates LLMs with external data sources for better insights

## Tutors

Provides personalized learning experiences using LLMs

## Chatbots with Memory

Enhances user interaction with persistent conversations

## Research Tools

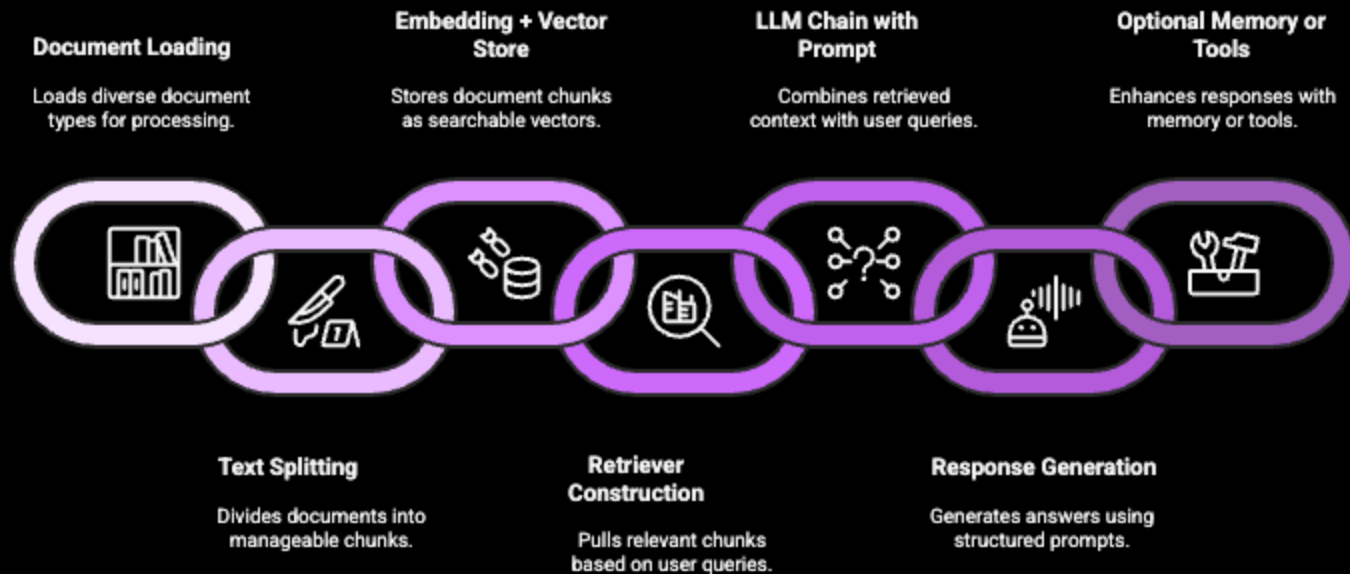
Facilitates research by analyzing and synthesizing information

**LangChain's  
Versatile  
Applications**



# LANGCHAIN

## LangChain RAG System Components



# OPENAI API

## OpenAI API Models Overview

### GPT-4

More capable and context-aware model



### GPT-4o

Fast, multimodal, and cost-effective model

### GPT-3.5

Fast and low-cost language model

## Capabilities of OpenAI Models

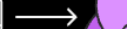
Natural Language Understanding



Natural Language Generation



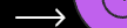
Code Generation



Summarization



Question Answering



Translation



Chat-based Applications



Made with  Napkin

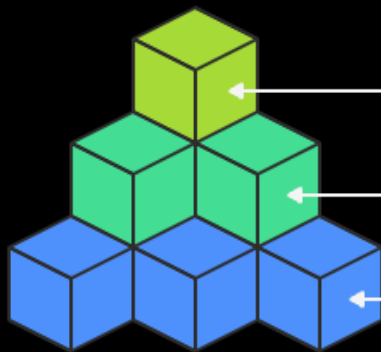
Made with  Napkin





# HUGGINGFACE

## Hugging Face's AI Ecosystem



### ML Collaboration Platform


Sharing and collaborating on ML resources

### Transformers Library

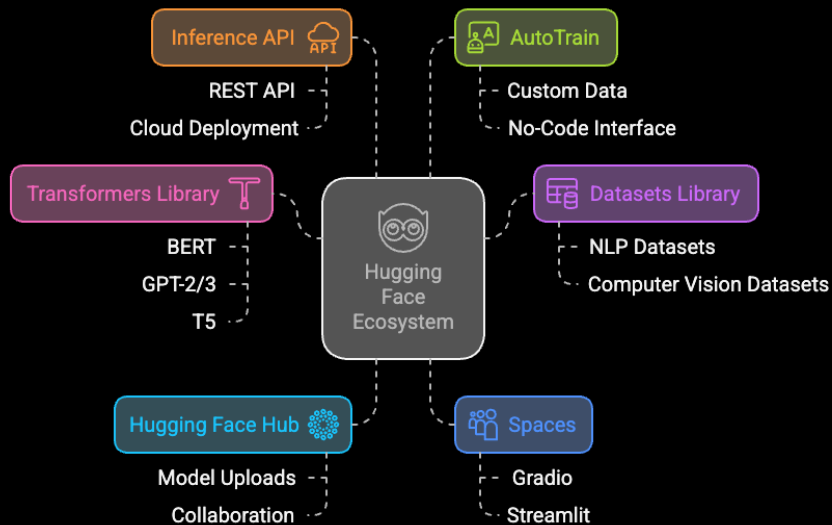
Provides pre-trained models for NLP

### Open-Source AI Hub

Platform for AI research and deployment

Made with  Napkin

## Hugging Face Ecosystem: Components and Functionalities



Made with  Napkin





# HANDS-ON





# **CODING CHALLENGE**





# **SOLUTION PRESENTATIONS**

