



A GenAI Agent for Statistical Output Review

Dr. Lizhong Liu

Agenda

- 01** Objectives
- 02** Obstacles
- 03** Key terms
- 04** Methodology
- 05** Future works
- 06** Q&A Session



UP NEXT

Objectives

Objectives

We aim to build an AI agent system that enhances the efficiency of clinical report analysis for Data Monitoring Committees (DMCs).



- Provide DMCs with real-time response to questions towards clinical data.
- Reduce the workload and time required to review extensive clinical documents.
- Enhance accuracy and reduce human error in data querying and analysis.



UP NEXT

Obstacles

Obstacles

Problems encountered:



Too Much Data to Review in short period



Delayed Responses for Ad-hoc request



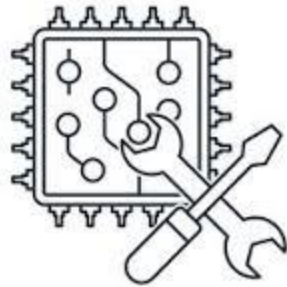
Unfriendly data presentation (PDF hard to process)

Generic LLM tech difficulties:

- **Hallucination:** Model's response are not reliable all times
- **Knowledge Out-of-date:** Model's knowledge are based on training data (property tax is different than current rate)
- **No external knowledge:** Model have no access to external or domain knowledges (no information about weather)

Obstacles

How does industries use LLM on specific aims?



Finetune model
(retrain model on specific task)

Resource Intensive



Retrieval-Augmented Generation
(Ask AI to use knowledge base to answer question)

Low Cost + Agile Dev



UP NEXT

Key terms



Key terms

- **Hallucination**
- **Embedding**
- **Vector Store**
- **RAG (Retrieval-Augmented Generation)**
- **ReAct Agent (Reasoning and Acting Agent)**

Key terms

- **Hallucination:** LLM generate seemingly factual but false or misleading information, since LLMs are based on recursive generation of the next token using probabilistic predictions of plausible continuations.



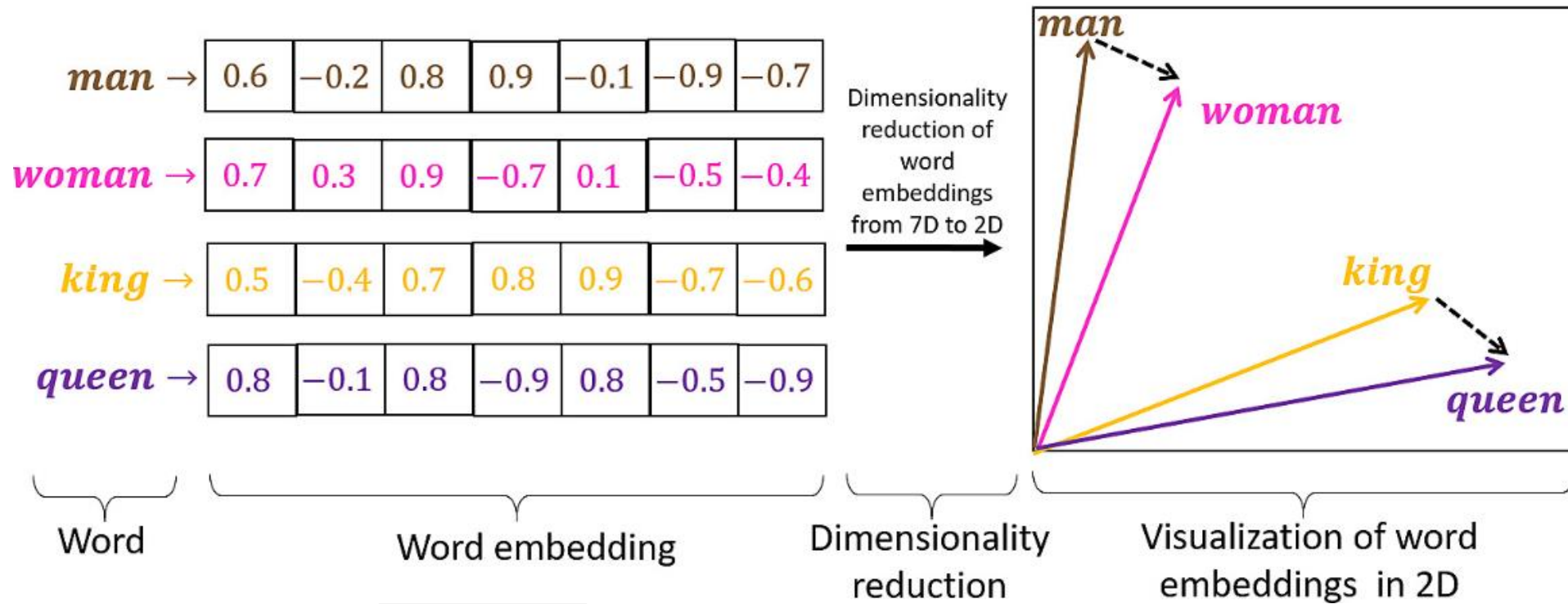
which one is bigger? 8.11 and 8.3



The number 8.3 is actually smaller than 8.11. When comparing decimal numbers, you look at the digits from left to right. Here, both numbers start with 8, but the next digit in 8.3 is 3, which is less than the 1 in 8.11. If you extend 8.3 to match the number of decimal places in 8.11, it becomes 8.30, which is still smaller than 8.11.

Key terms

- **Embedding:** A mathematical representation of data (like words or documents) in a high-dimensional vector space, capturing semantic relationships



Key terms

- **VectorStore:** A specialized database that stores vector embeddings, allowing for fast similarity searches and data retrieval.

Vector Stores

1. Load Source Data



Load, Transform, Embed

Vector Store



2. Query Vector Store

Embed

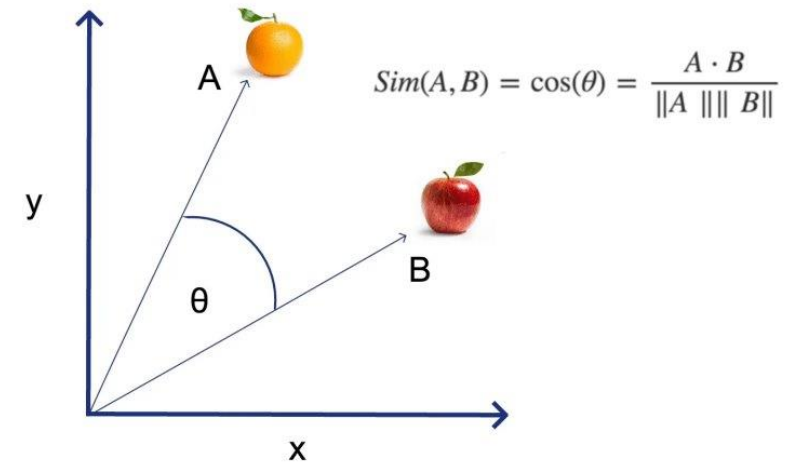
5.5, -0.3...
2.1, 0.1

XXXXXXXXXXXXX
XXXXXXXXXXXXX

XXXXXXXXXXXXX
XXXXXXXXXXXXX

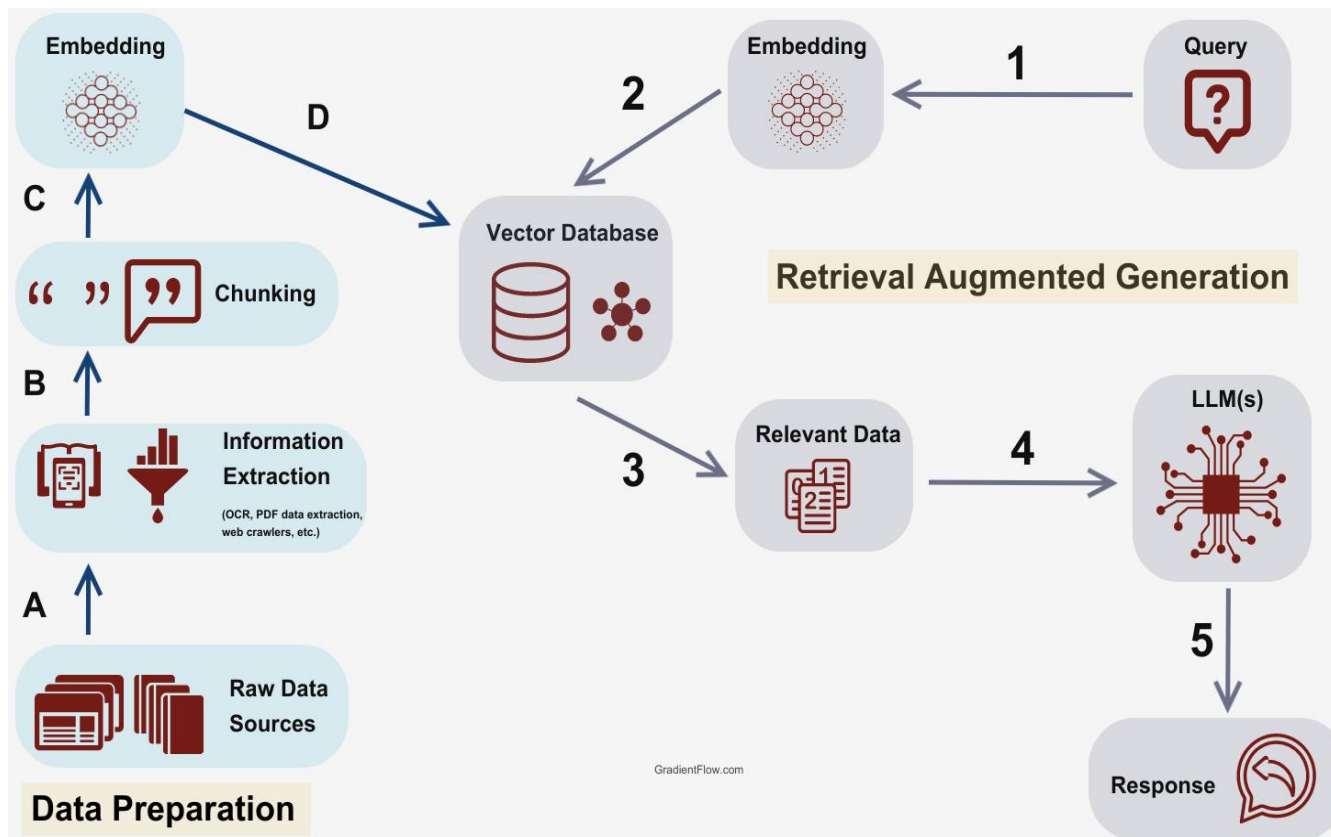
3. Retrieve 'most similar'

Cosine Similarity

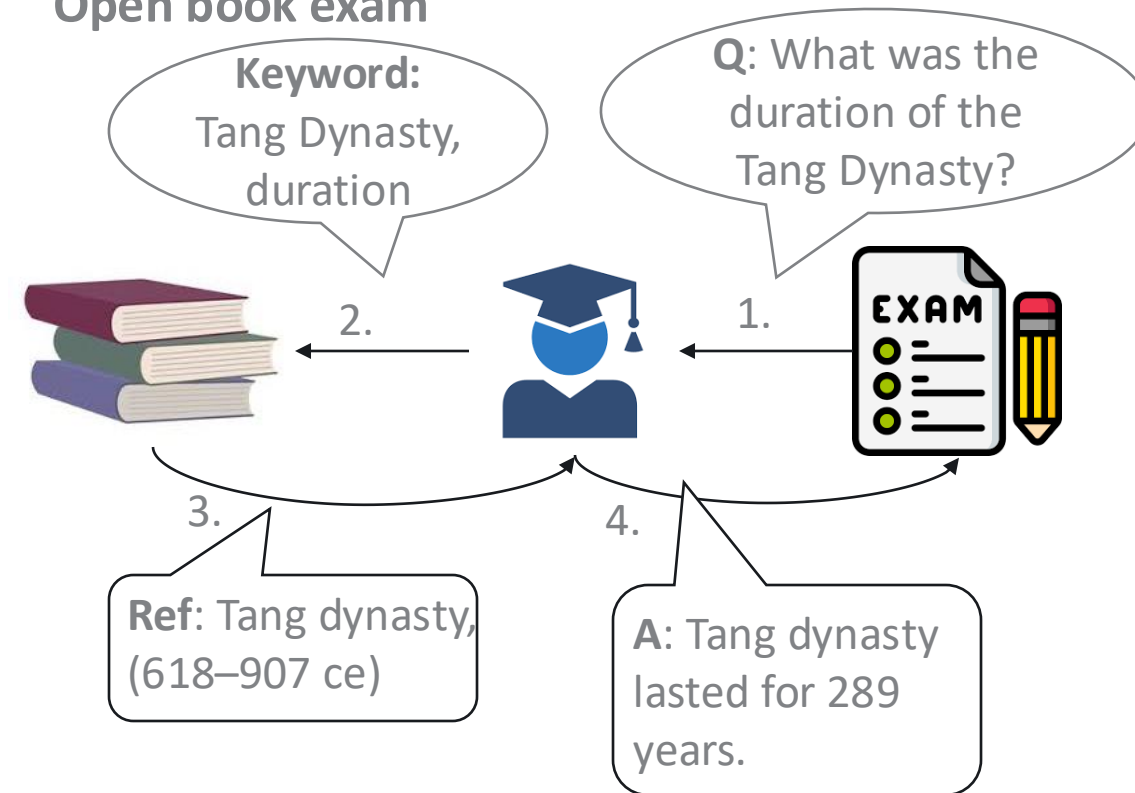


Key terms

- **RAG (Retrieval-Augmented Generation):** A framework that combines retrieval of relevant information from knowledge base with a generative language model to produce accurate and contextually relevant responses



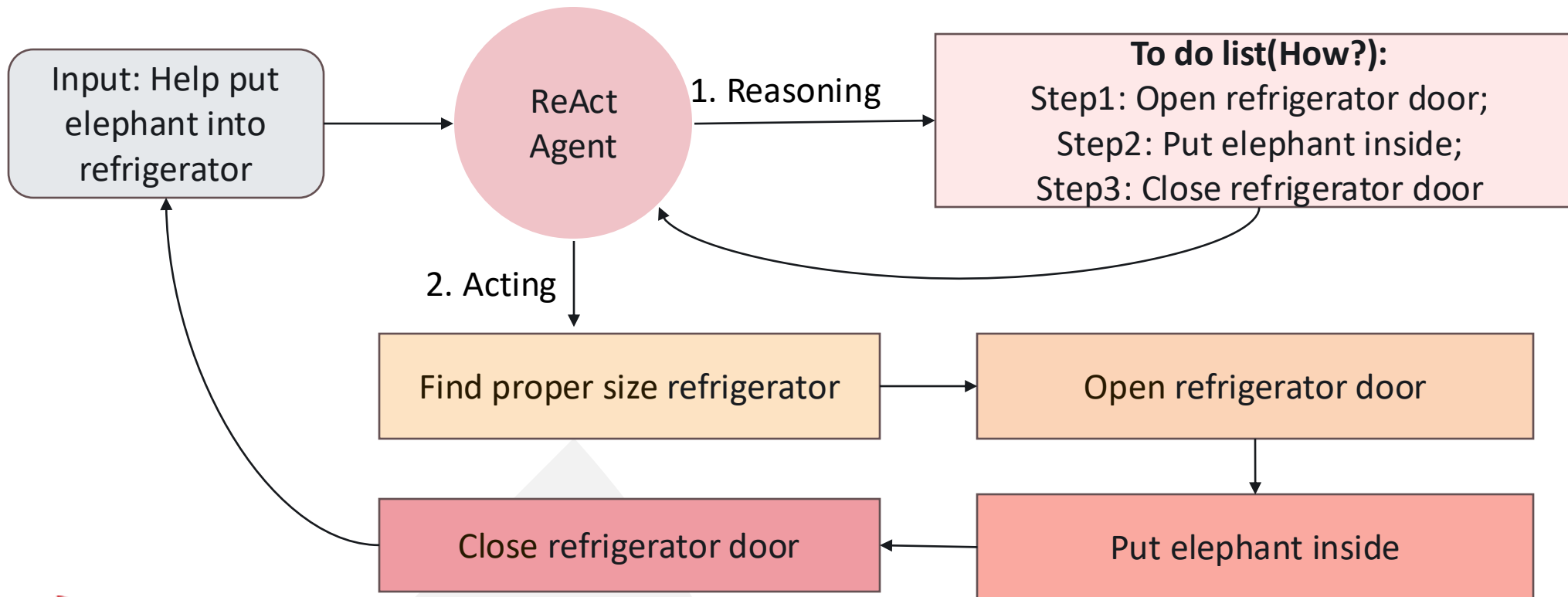
Open book exam



Without reference, LLM is taking closed book exam, rely on knowledge learned during training period.

Key terms

- **ReAct Agent(Reasoning and Acting Agent):** An autonomous entity observes its environment, makes decisions, and performs actions to achieve specific goals by combining reasoning (thought processes) and acting (executing tasks) capabilities to solve problems effectively.





UP NEXT

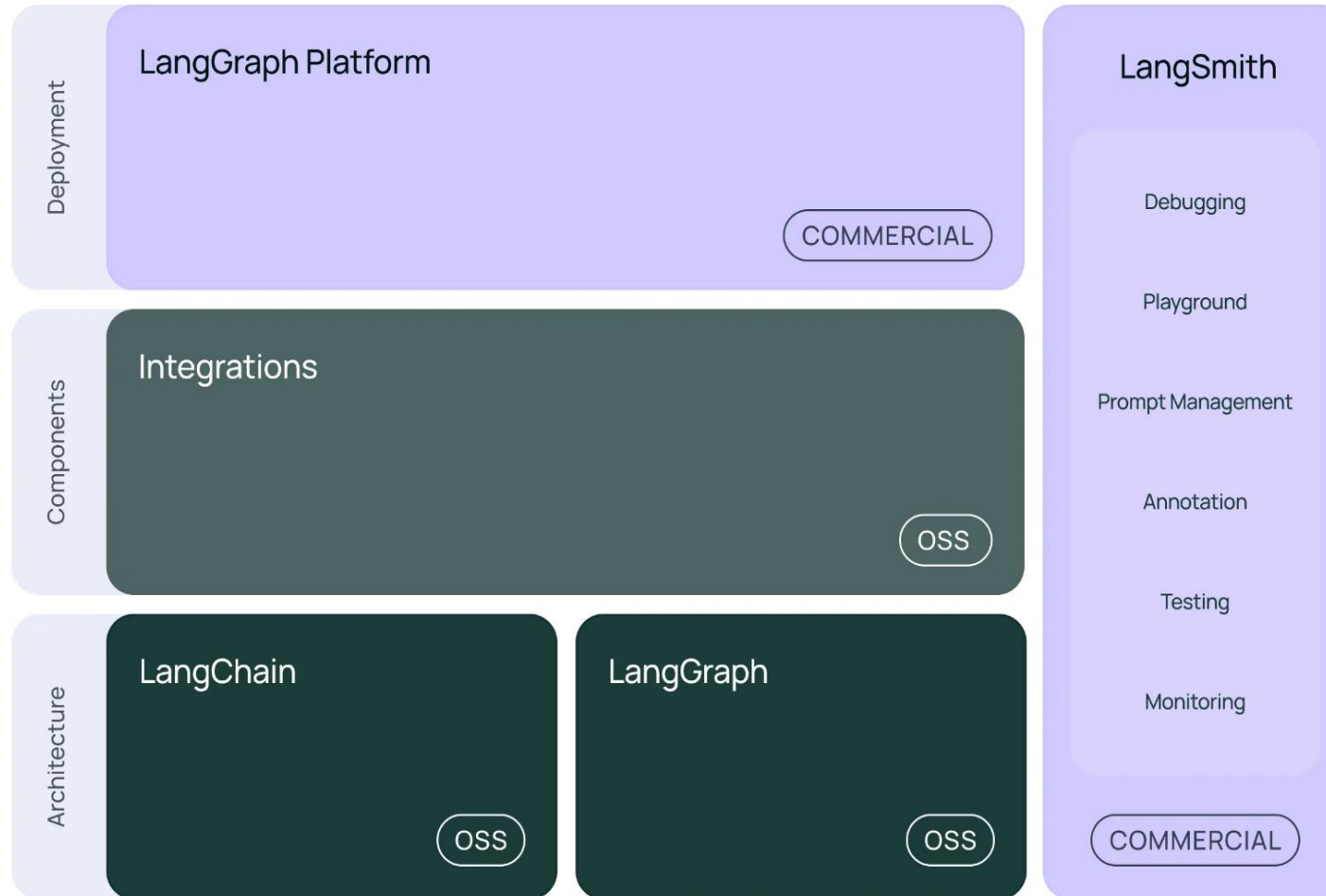
Methodology

Methodology

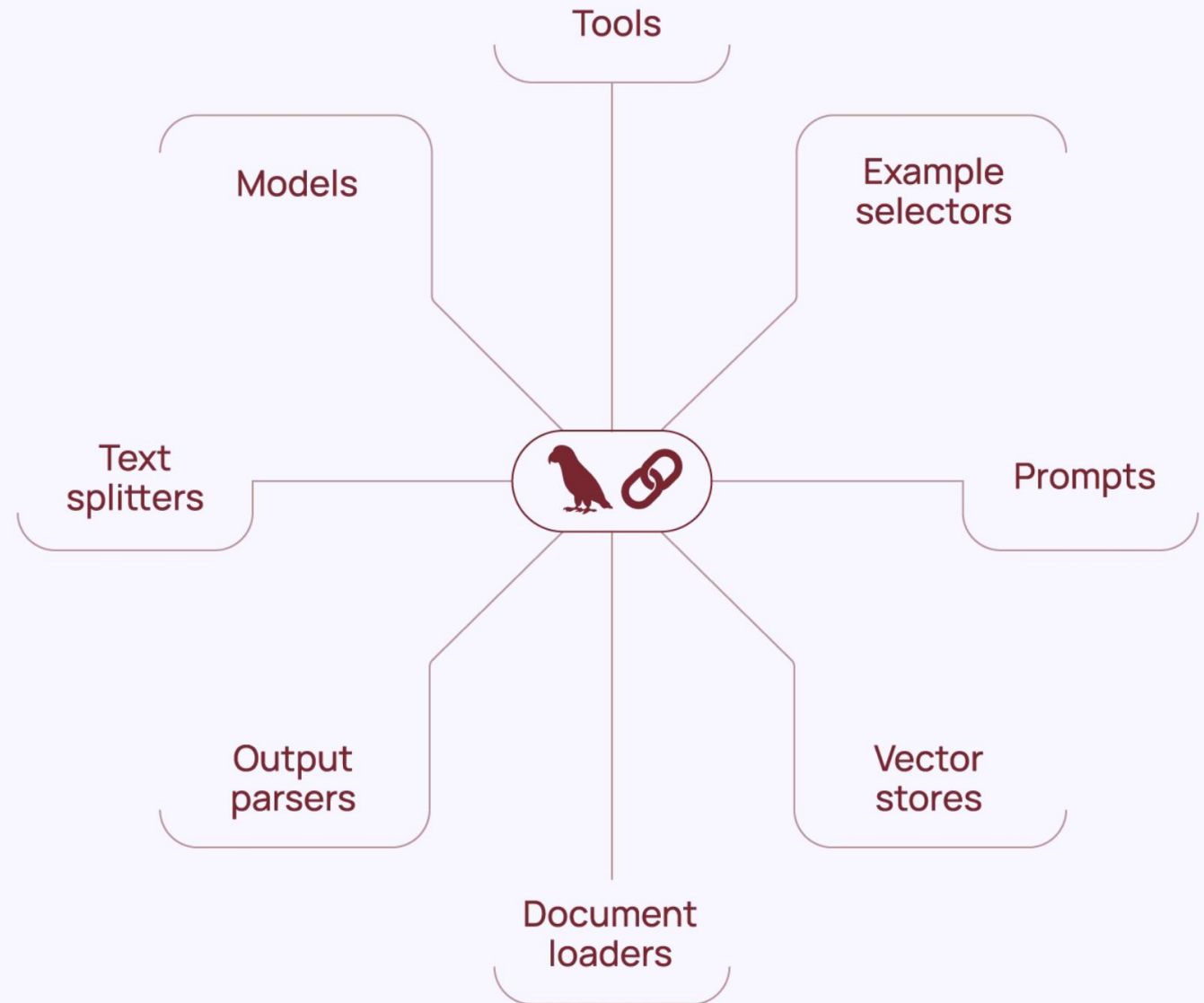
Popular AI develop framework



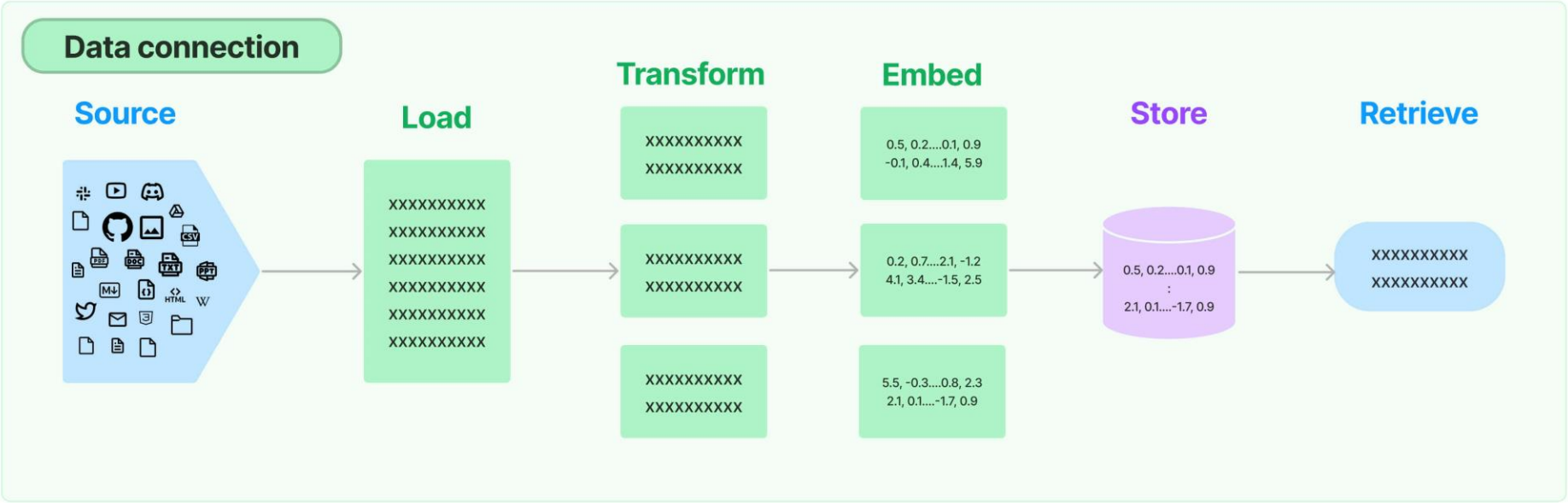
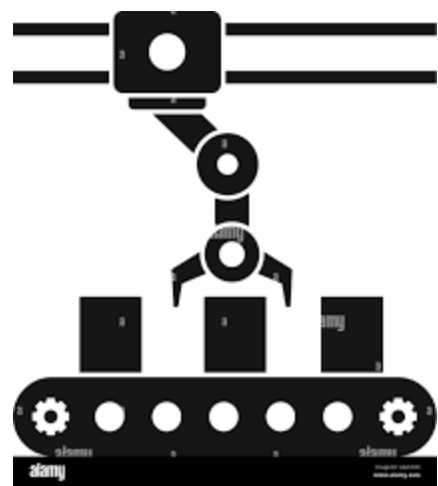
LangChain is considered as most robust
AI develop framework for now



LangChain

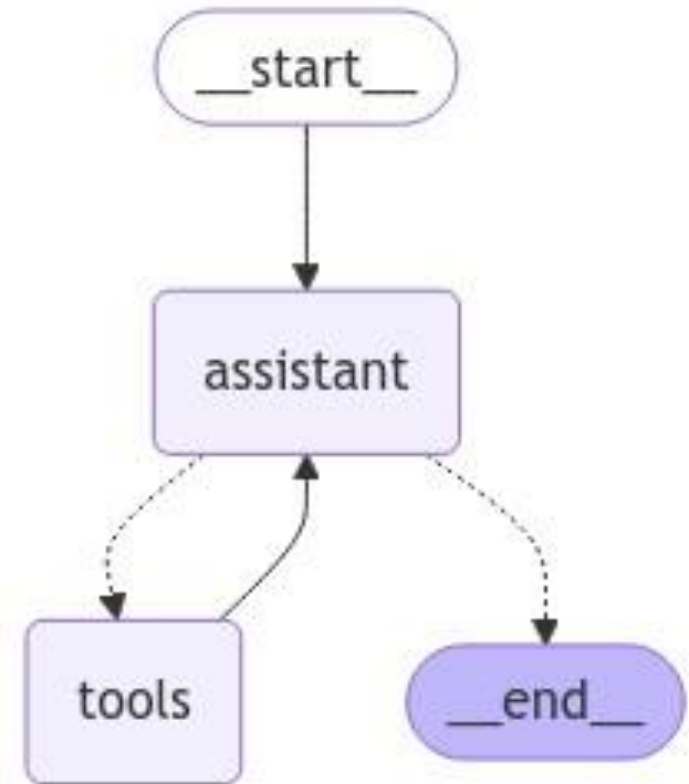
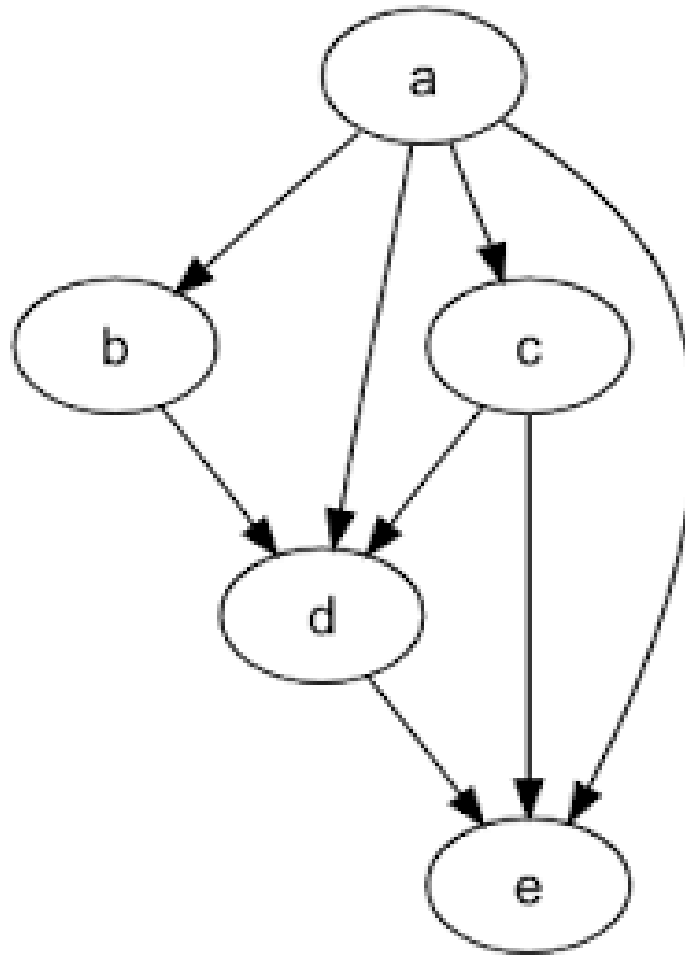


LangChain allow you to build your applications using LangChain's open-source building blocks, components, and third-party integrations



You can choose proper modules and combine them to form a chain that execute like assembly line.

LangGraph use DAG(directed acyclic graph) to design more complex workflow that enables human interaction, conditional edge and multi-agents interactions.





I am DARIA

Your Data Analysis & Review Intelligent Assistant



Upload data (PDF,
CSV, SAS7BDAT,
RTF)



Ask questions on your
data

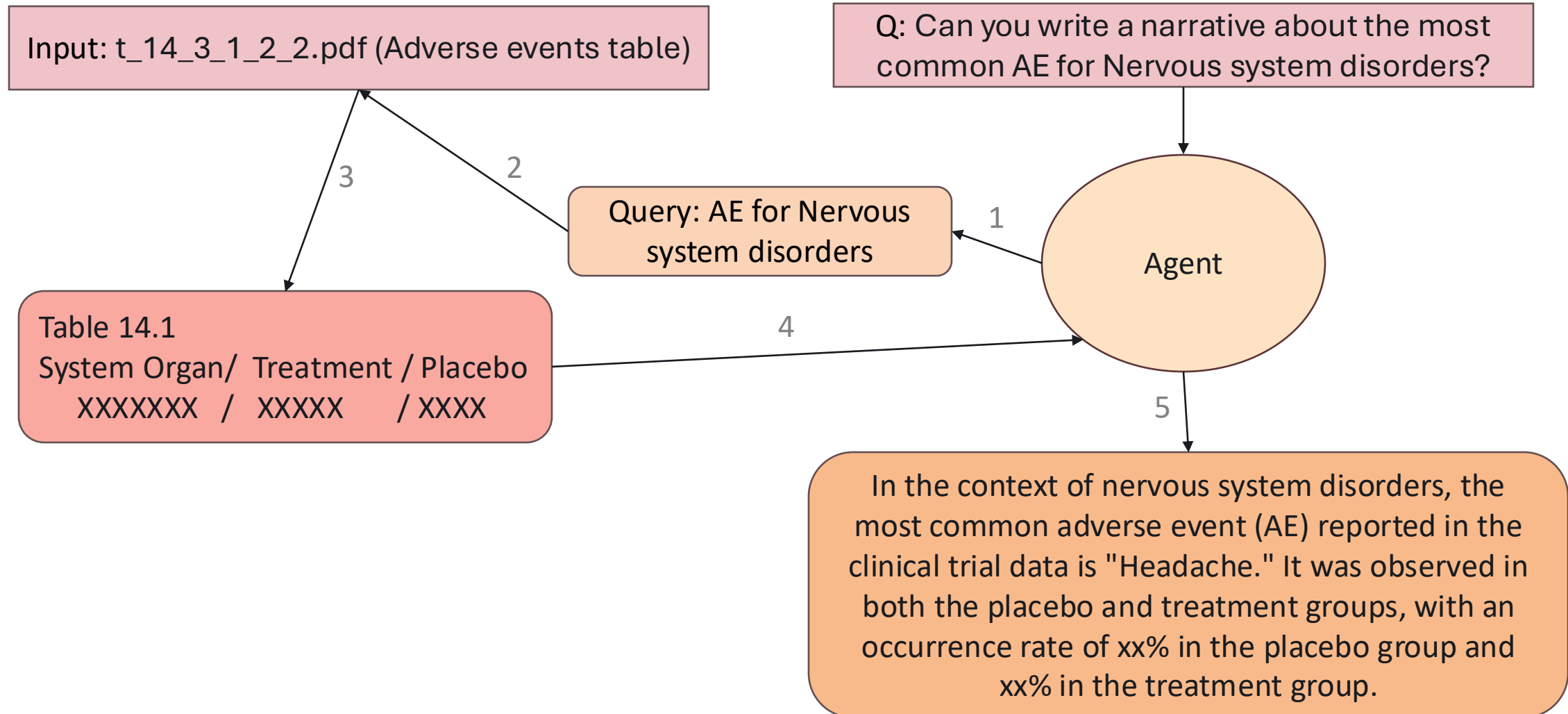


Generate Charts and
Graphs



Download chat
session

DARIA Examples



Data Analysis and Review Intelligent Assistant (DARIA)

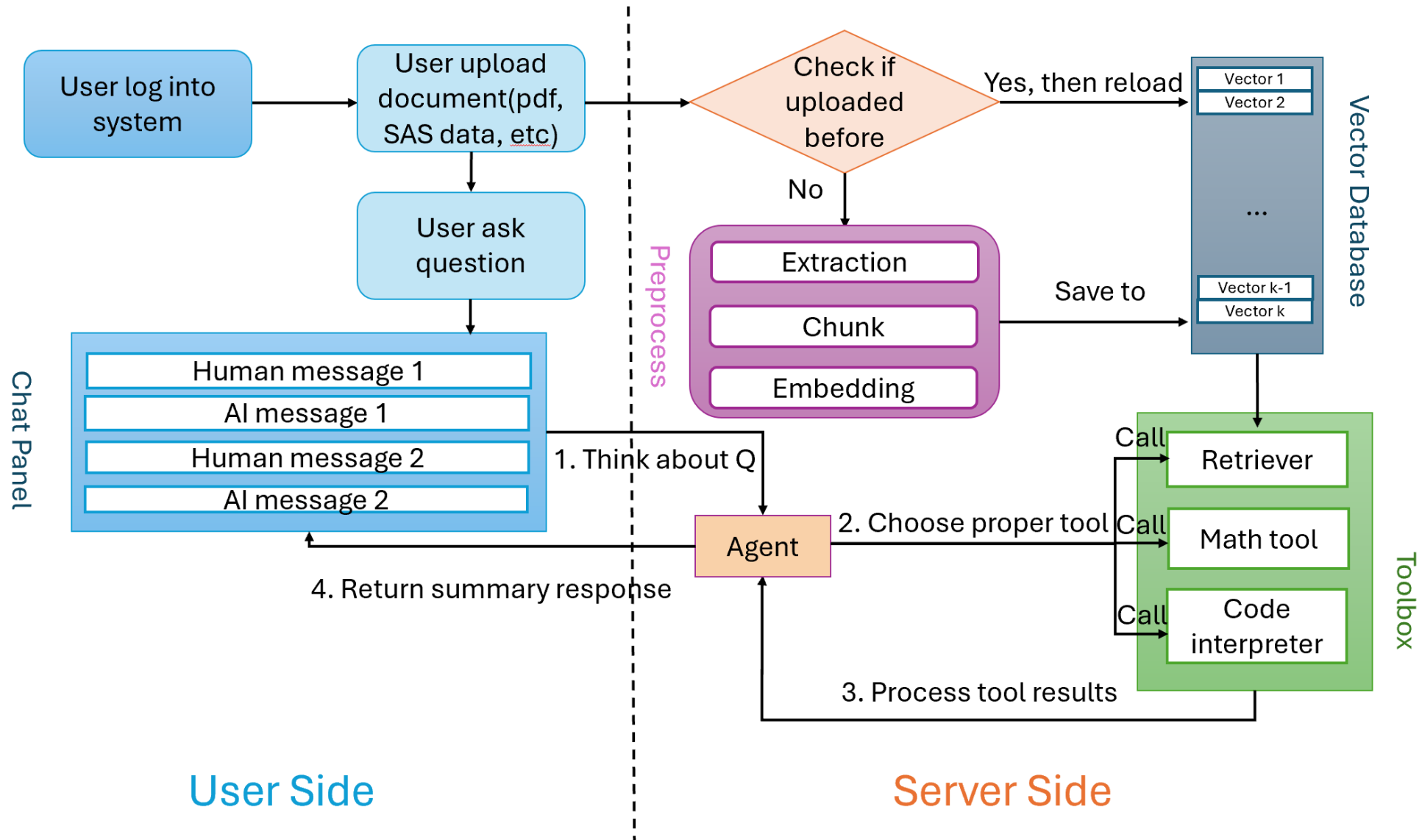


Figure 2. Structure design of DARIA system

Without RAG, for generic LLM, it is limited by token length:

⚠ This file contains too much text content. Please try again with a smaller file.



With our RAG system, since Vector Store are localized and we only query K/N results; We can upload and query file without limitation.

Listing 16.2.3 Analysis populations and reasons for exclusion - Randomized patients

Listing 16.2.4.1 Demographics - Randomized patients

1

23468

Pros

Flexibility: tools are extensible and can be customized for different applications. LLMs can be easily changed. Framework can be easily converted for different needs.

Scalability: supports modular development, enabling individual components to be developed and combined seamlessly.

Cons

Steep Learning Curve: requires a solid foundation in several areas, including prompt engineering, Python object-oriented programming, graph theory and even APIs and hardware knowledges, which can be challenging for new users.

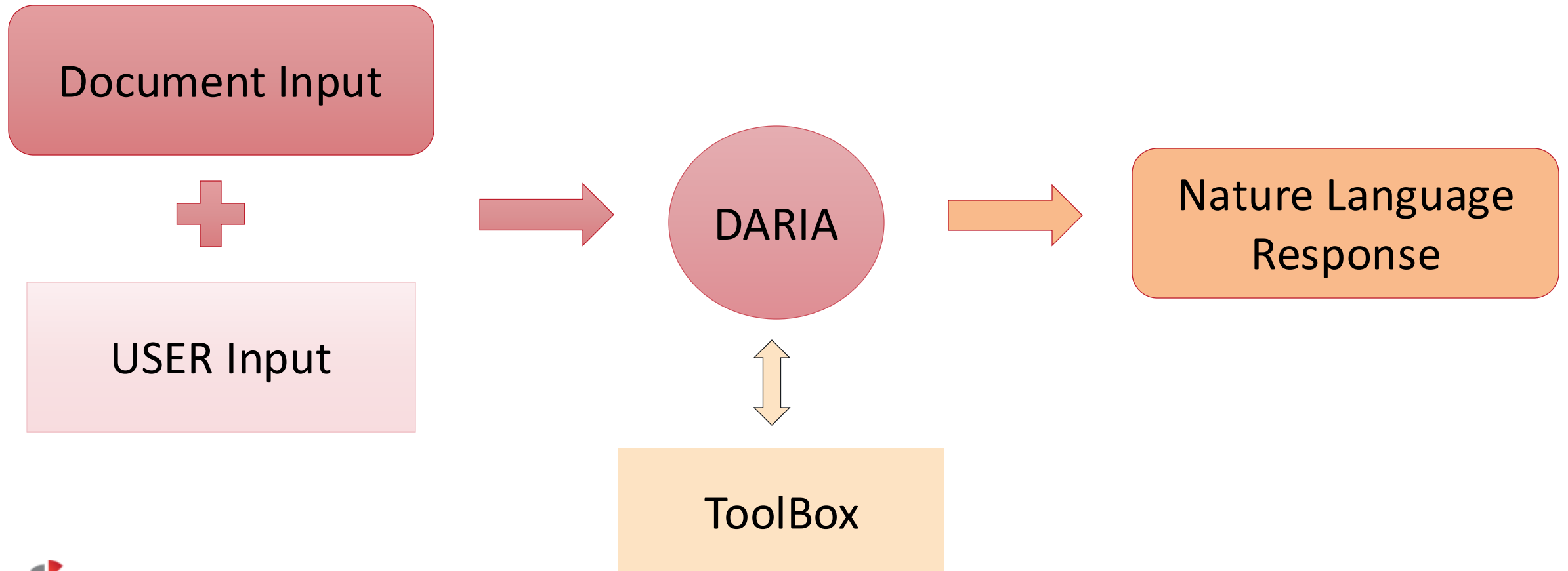
Limited Documentation and Resources: Limited documentation and resources, along with frequent system updates, make replication and debugging challenging and require continuous time investment for adapting to changes.



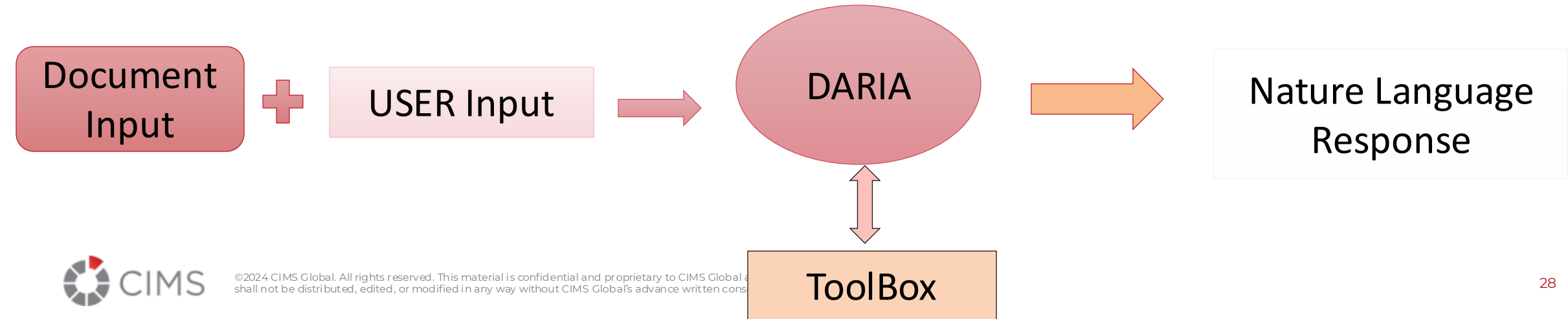
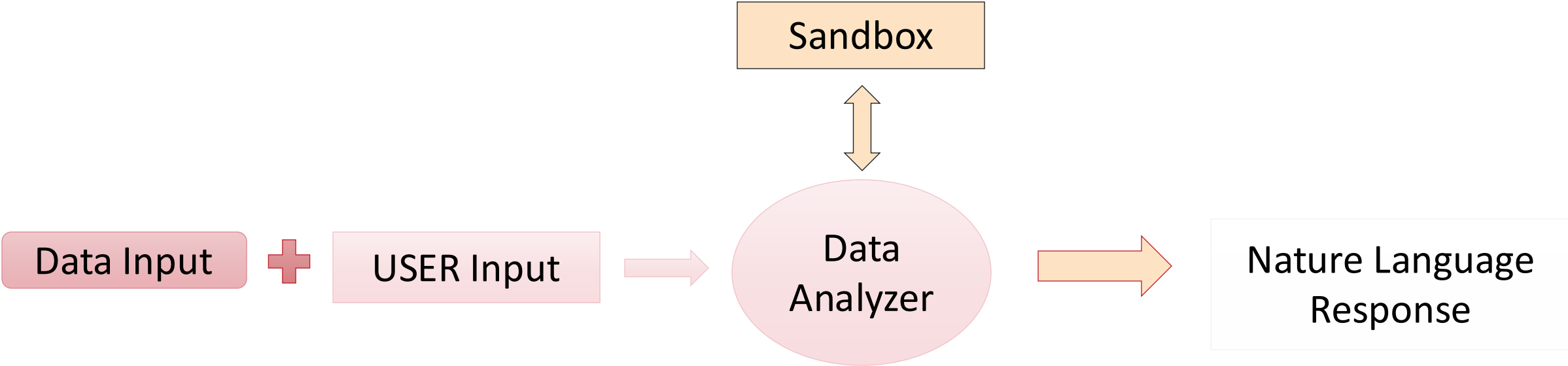
UP NEXT

Future works

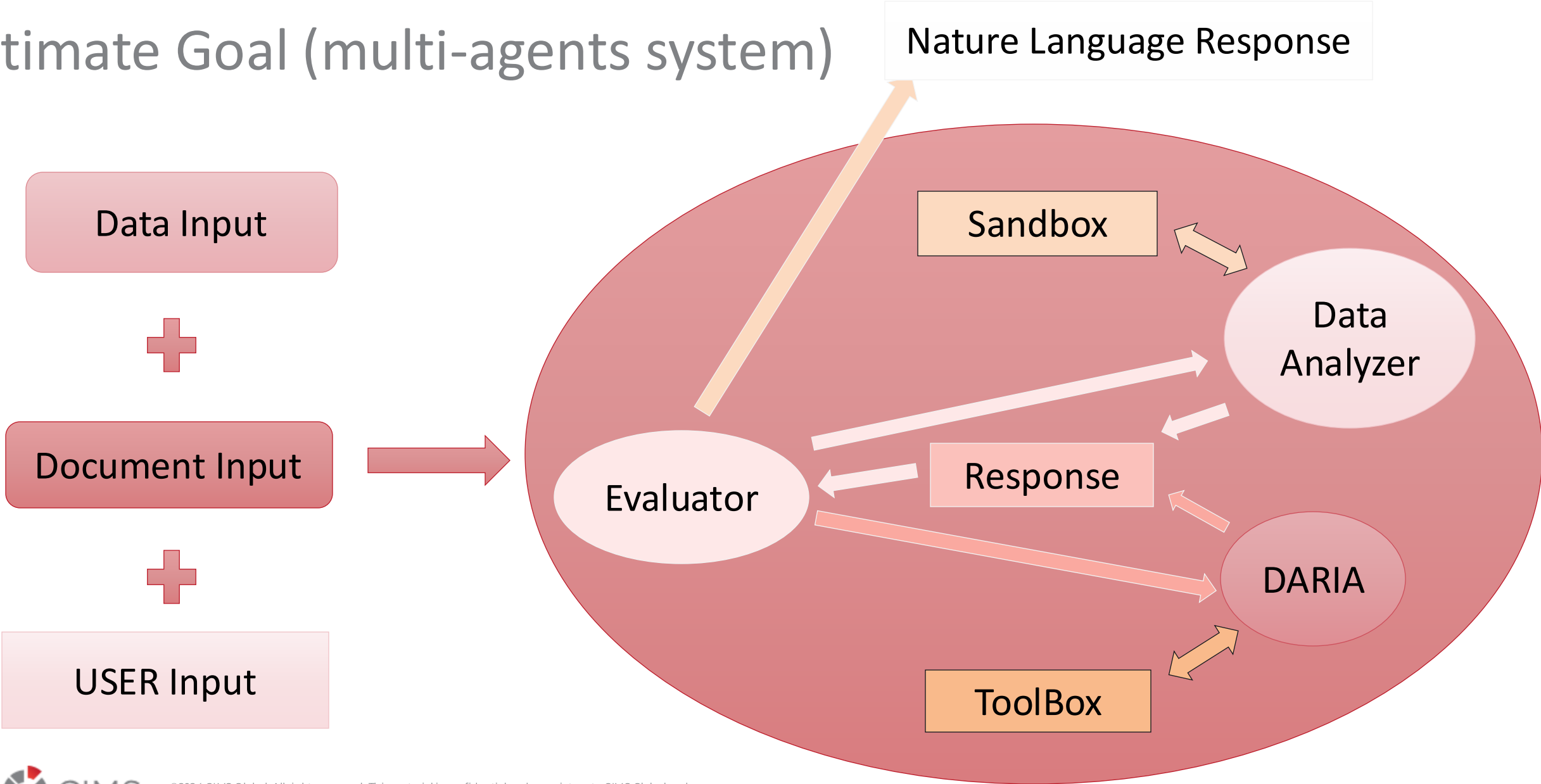
Current Design



Current Solution



Ultimate Goal (multi-agents system)



Agent Community

Q&A Session

Please submit your questions to the moderator.

Thank you.

We welcome any questions or thoughts that you may have.

john.doe@cims-global.com / 908.557.9366



©2024 CIMS Global. All rights reserved. This material is confidential and proprietary to CIMS Global and shall not be distributed, edited, or modified in any way without CIMS Global's advance written consent.