



# Pathway Analysis Assistant: PathwayGenie

Bett Hanchu Zhang

2024.12.23

## Table of Contents

- 1. Introduction ..... 2
- 2. Data Modeling ..... 2
- 3. Agent Architecture ..... 4
- 4. System Architecture ..... 6
- 5. Tools and Technologies: ..... 8
- 6. Examples ..... 8
  - Case1:"Find the diseases associated with Gene 'MAPK8'." ..... 9
  - Case2:Find the diseases associated with Gene 'SOCS4' , and find other entities related with this disease. .... 9
  - Case3:Analyze the downstream effects of 'GCK' on Type II diabetes mellitus. .... 11
- 7. System Analyze ..... 13
  - 7.1 Strengths ..... 13
  - 7.2 Future Enhancements ..... 13
- 8. References ..... 14

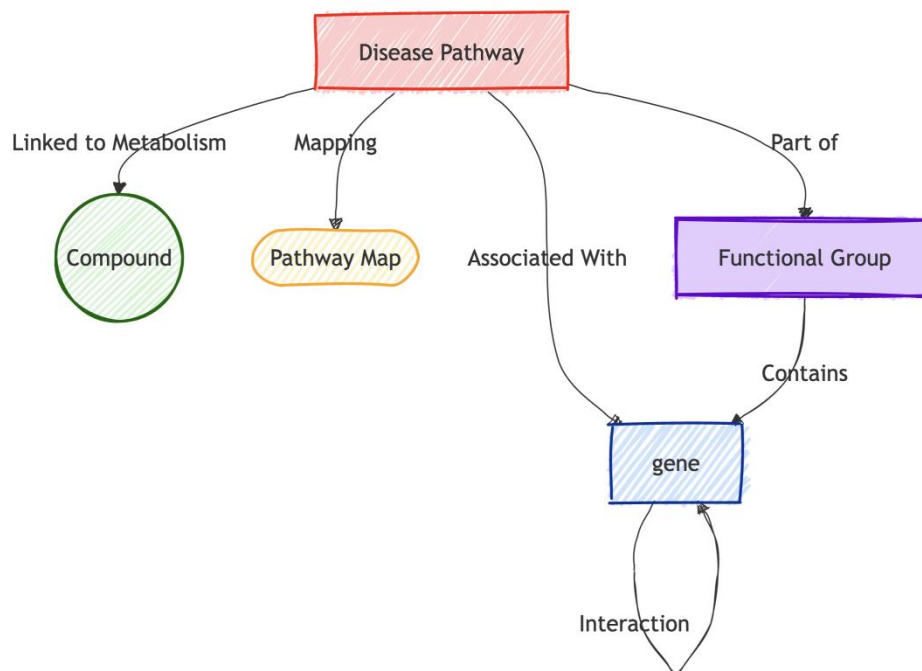


# 1. Introduction

PathwayGenie is a biomedical assistant based on agent architecture, which designed to help researchers uncover the complex interactions between genes, pathways, and diseases. Leveraging data from KEGG and Gene Ontology (GO), it provides insights into molecular processes, identifies gene–disease associations, and performs downstream analysis to generate actionable hypotheses.

## 2. Data Modeling

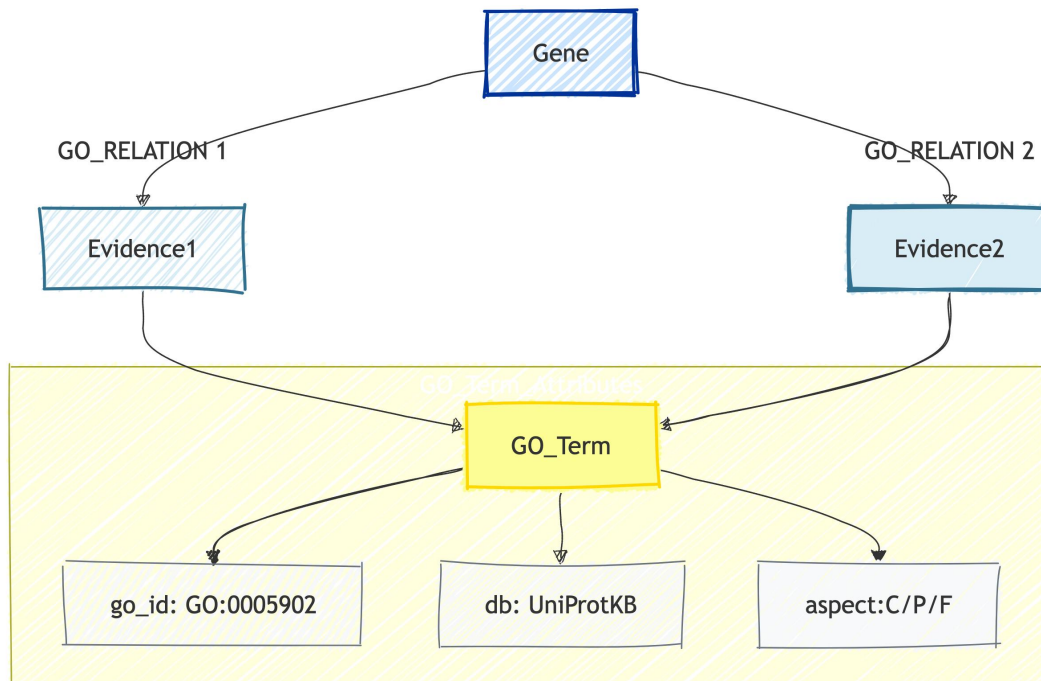
The schema is designed to capture the relationships between key components extracted from KGML files. These files contain data about Diseases, as well as entities such as Genes, Compounds, Maps, and Group, along with the relationships between them.



*Picture 1 entries and relationships in KGML files*

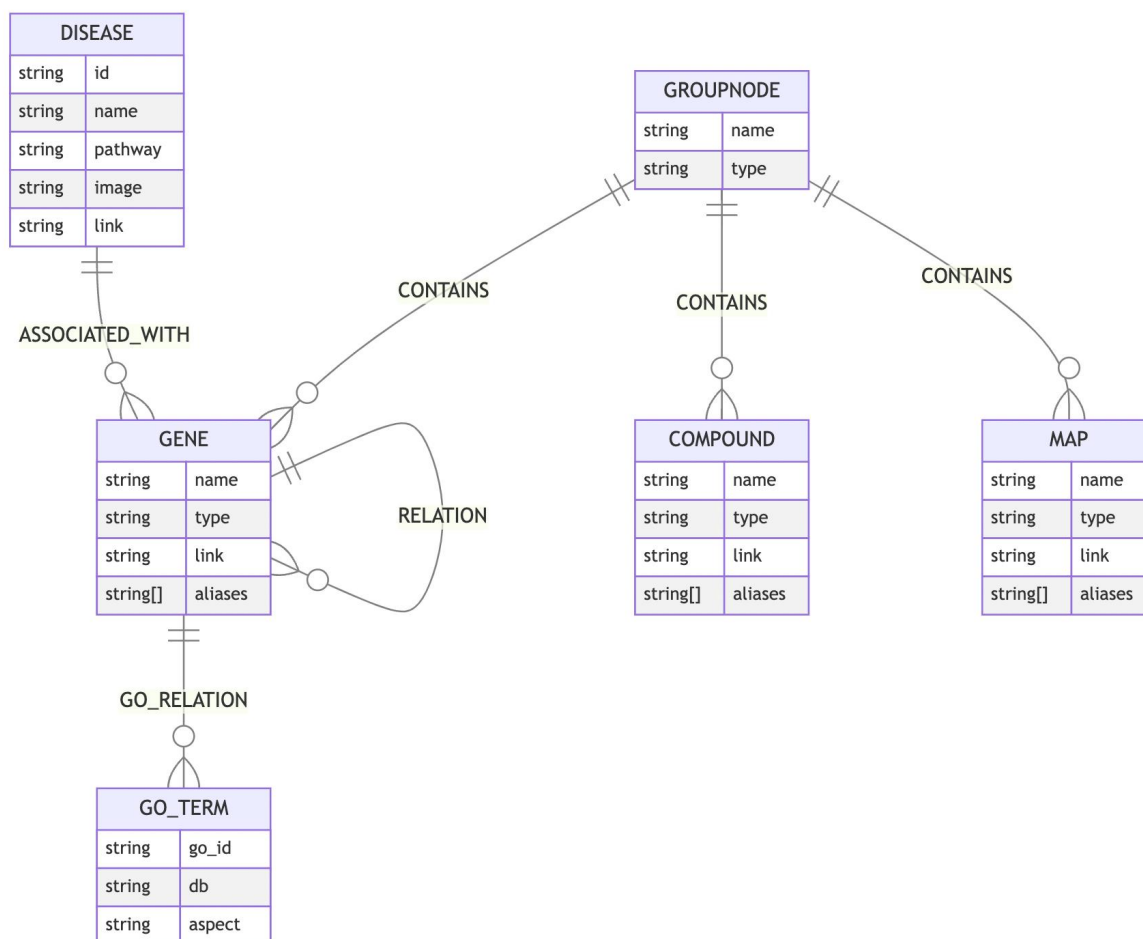


Gene Ontology(GO) files of contains gene-to-GO relationship. GO\_Term with attributes like go\_id, aspect, and description. Their relationship also keeps track of evidence\_code, assigned\_by, date, some infomation.



*Picture 2 Gene-GO Relationship*

Integrate data by linking diseases, genes, and GO terms with their relationships, then construct a graph structure in Neo4j.

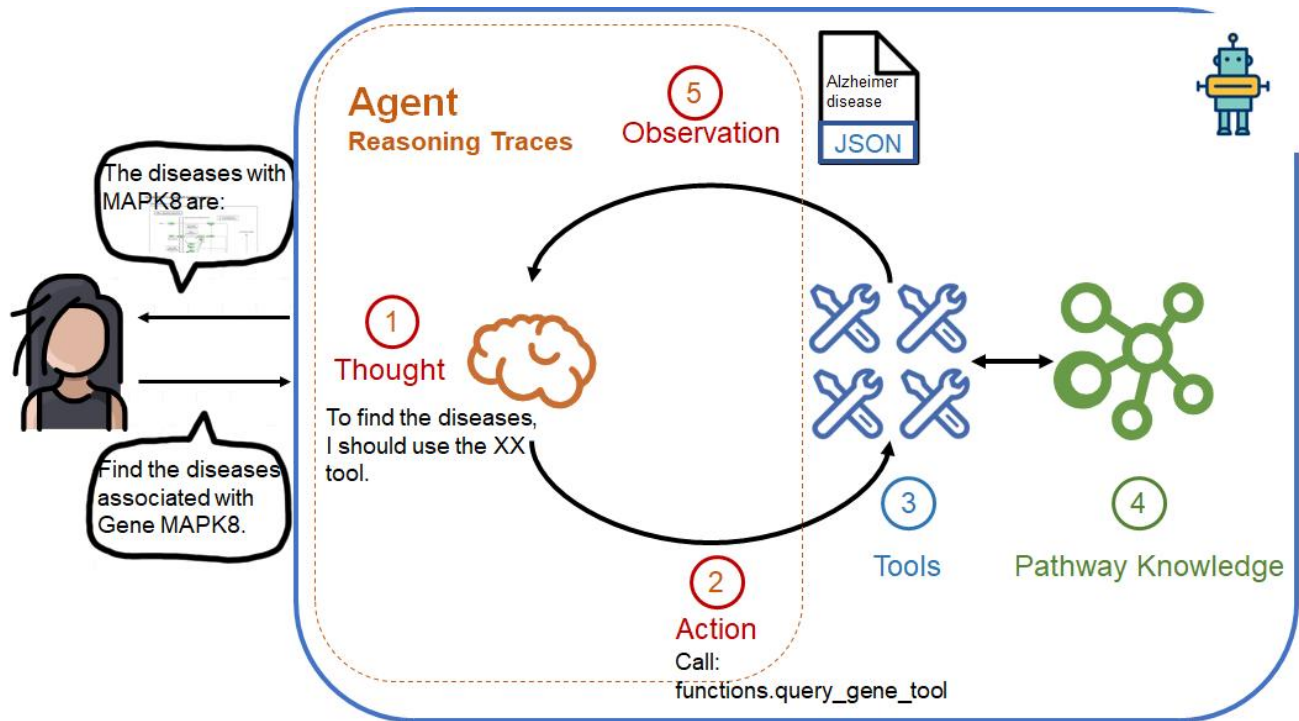


Picture 3 Data Structure in Neo4j

### 3. Agent Architecture

The ReAct[3] pattern, short for “Reasoning and Acting,” is a framework that mimic humans critical thinking behavior, combining thinking, action, observation steps.

At its core, the ReAct pattern works, the "thoughts" allow the LLMs to strategize and track progress (task solving trajectories), while the actions let them actually accomplish goals by using tools. By feeding observations to the LLM, allowing it to re-evaluate all information from trajectories and acts on these insights to make a informed decision making for the next attempt. Thus improve its reasoning capabilities.

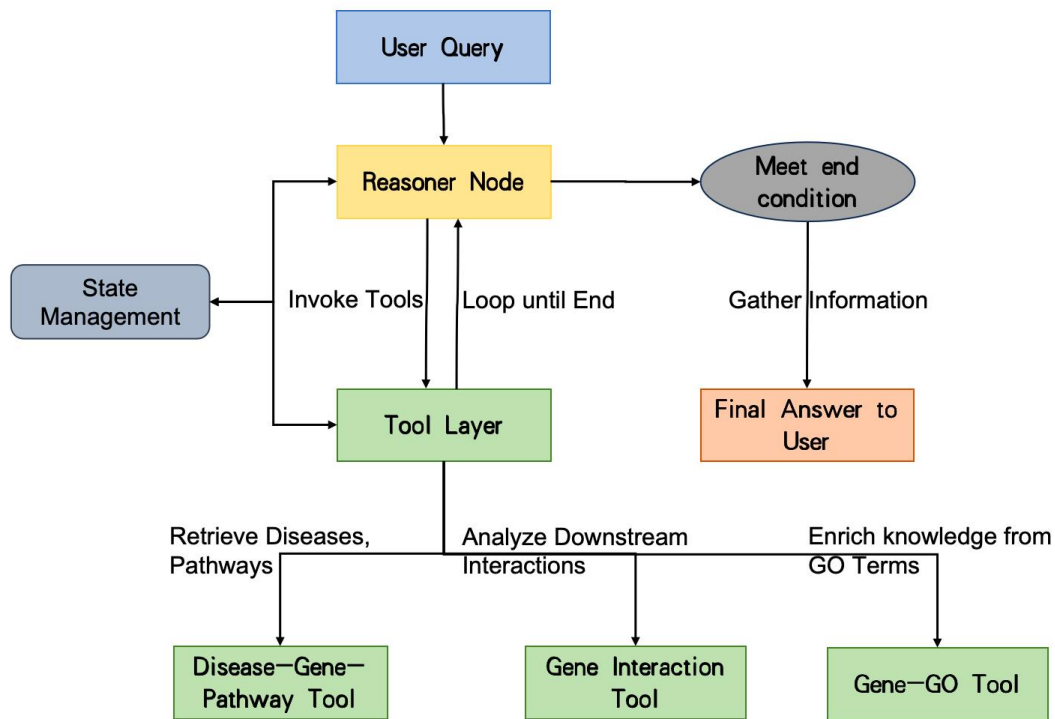


Picture 4 ReAct Agent Thought+Action+Observation Loop based on Pathway KG

In the context of pathway analysis, the ReAct pattern is particularly valuable for handling complex queries involving genetic pathways, diseases, and cascade downstream interactions.



## 4. System Architecture



*Picture 5 System Architecture*

### User Query

The entry point of the workflow where users input their queries.

### Reasoner Node

The central decision-making component of the workflow.

Responsible for:

- Interpreting the user query.
- Breaking it into actionable sub-tasks.
- Invoking appropriate tools to gather the required information.
- Compiling intermediate results into a final answer.



Operates iteratively, maintaining state information across reasoning steps.

## Tool Layer

Composed of specialized tools for querying and analysis:

- `query_gene_diseases_pathway_tool`: Retrieves diseases, pathways associated with a given gene.
- `query_gene_interaction_downstream_tool`: Analyzes downstream interactions of a gene and predicts their potential effects on pathways and diseases.
- `query_gene_diseases_pathway_tool`: Retrieves gene–GO term.

Tools are conditionally invoked based on the Reasoner Node's logic and query requirements.

## State Management

Managed by `GraphState`, which tracks:

- Messages exchanged between nodes (messages).
- The current reasoning step (`current_turn`).

Ensures seamless execution of the workflow by preserving intermediate states and transitions.

## Tool Layer

If a tool is required, the workflow transitions to the Tool Layer.

After tool execution, the state is updated.

Reasoner will do the observation tool action result, based on the results do the thought and then decision–making, decide whether to use tools for the next step, acquire more information from KG. Reasoning, observation, Action, continues until a complete answer is generated.



## End Condition

If the Reasoner Node calls a tool, the workflow routes to the Tool Layer.

If no tool is required, the workflow concludes and generates the final answer.

## Output Generation

Integrates all gathered information from the tools and reasoning steps.

Produces a detailed and comprehensive report as the final answer to the user query.

# 5. Tools and Technologies

List tools and used: Neo4j, GPT4, LangGraph/LangChain

Agent frameworks: LangGraph/LangChain

Agent debug tool: LangSmith

Deploy and User Interface: LangGraph Platform

# 6. Examples

Example queries include:





- Case1:"Find the diseases associated with Gene 'MAPK8'."

```
The gene 'MAPK8' is associated with the following diseases:

1. [Type II diabetes mellitus](https://www.kegg.jp/kegg-bin/show_pathway?hsa04930)
! [Type II diabetes mellitus](https://www.kegg.jp/kegg/pathway/hsa/hsa04930.png)

2. [Alzheimer disease](https://www.kegg.jp/kegg-bin/show_pathway?hsa05010)
! [Alzheimer disease](https://www.kegg.jp/kegg/pathway/hsa/hsa05010.png)

3. [Parkinson disease](https://www.kegg.jp/kegg-bin/show_pathway?hsa05012)
! [Parkinson disease](https://www.kegg.jp/kegg/pathway/hsa/hsa05012.png)

4. [Colorectal cancer](https://www.kegg.jp/kegg-bin/show_pathway?hsa05210)
! [Colorectal cancer](https://www.kegg.jp/kegg/pathway/hsa/hsa05210.png)

These diseases are connected to the 'MAPK8' gene in different pathways.
```

Picture 6 Generated Report

- Case2:Find the diseases associated with Gene 'SOCS4' , and find other entities related with this disease.

Thought 1: To find out the diseases associated with the gene 'SOCS4', I can use the `query_gene_diseases_pathway_go_tool`. After finding the diseases, I can use the `query_gene_interaction_downstream` tool to find other entities related to this disease.

Action1: Function1 Calling

Observation1: The gene 'GCK' is associated with the disease 'Type II diabetes mellitus'. You can find more details about this disease [here](https://www.kegg.jp/kegg-bin/show\_pathway?hsa04930)

Thought2: Now, let's find other entities related to 'Type II diabetes mellitus'.

Action2: Function2 Calling

Observation2: `related_entity: name: hsa:5313 hsa:5315 link: https://www.kegg.jp/dbget-bin/www_bget?hsa:5313+hsa:5315 type: Gene aliases: PKLR`



*Final Answer:*

The 'SOCS4' gene is associated with the disease 'Type II diabetes mellitus'. The entities that are directly related to 'Type II diabetes mellitus' include the following genes:

1. PKLR, PK1, PKL, PKRL, RPK...
2. P3R3URF-PIK3R3, LOC110117498-PIK3R3...
3. SLC2A4, GLUT4
4. IRS1, HIRS-1
5. INSR, CD220, HHF5
6. INS, IDDM, IDDM1, IDDM2, ILPR, IRDN, MODY10, PNDM4
7. SOCS4, SOCS7...
8. MAPK1, ERK, ERK-2, ERK2, ERT1, MAPK2, NS13, P42MAPK, PRKM1, PRKM2, p38, p40, p41, p41mapk, p42-MAPK...
9. ADIPOQ, ACDC, ACRP30, ADIPQTL1, ADPN, APM-1, APM1, GBP28
10. IKBKB, IKK-2, IKK-beta, IKK2, IKKB, IMD15, IMD15A, IMD15B, NFKB1KB

*Picture 7 Generated Report*



● Case3:Analyze the downstream effects of 'GCK' on Type II diabetes mellitus.

Reflecion Steps	Thought	Action	Observation
1	To analyze the downstream effects of 'GCK' on Type II diabetes mellitus, we need to first <i>identify the diseases</i> associated with the gene 'GCK'.	functions calling	<ul style="list-style-type: none"><li>properties(d)<ul style="list-style-type: none"><li>Id 04930</li><li>Image <a href="https://www.kegg.jp/kegg/pathway/hsa/hsa04930.png">https://www.kegg.jp/kegg/pathway/hsa/hsa04930.png</a></li><li>Link <a href="https://www.kegg.jp/kegg-bin/show_pathway?hsa04930">https://www.kegg.jp/kegg-bin/show_pathway?hsa04930</a></li><li>Name Type II diabetes mellitus</li><li>Pathway path:hsa04930</li></ul></li></ul>
2	The gene 'GCK' is associated with the disease 'Type II diabetes mellitus'. Now, we need to <i>identify the entities that interact with 'GCK'</i> to understand its downstream effects.	functions calling	<ul style="list-style-type: none"><li>0<ul style="list-style-type: none"><li>Gene<ul style="list-style-type: none"><li>Link <a href="https://www.kegg.jp/dbget-bin/www_bget?...">https://www.kegg.jp/dbget-bin/www_bget?...</a></li><li>Name hsa:2645 hsa:3098 hsa:3099 hsa:3101 hs...</li><li>Type Gene</li><li>Aliases</li></ul></li><li>Related Entity<ul style="list-style-type: none"><li>Link <a href="https://www.kegg.jp/dbget-bin/www_bget?...">https://www.kegg.jp/dbget-bin/www_bget?...</a></li><li>Name hsa:5313 hsa:5315</li><li>Type Gene</li><li>Aliases</li></ul></li></ul></li></ul>
3	The gene 'GCK' interacts with the genes 'PKLR' and 'PK1'. To understand the biological processes, molecular functions, and cellular components associated with these genes, we <i>need to query the GO terms</i> for these genes.	functions calling	<div>Result <input type="checkbox"/></div> <ul style="list-style-type: none"><li>0<ul style="list-style-type: none"><li>properties(go)<ul style="list-style-type: none"><li>Aspect F</li><li>Db ComplexPortal</li><li>Go Id GO:0004743</li></ul></li><li>properties(r)<ul style="list-style-type: none"><li>Type enables</li></ul></li><li>Evidences<ul style="list-style-type: none"><li>0<ul style="list-style-type: none"><li>{"evidence_code": "IBA", "reference": "GO_REF:0000033", "date": "20240729", "assigned_by": "GO_Central", "gene_obj_type": "protein"}</li></ul></li></ul></li></ul></li></ul>



4	<p><i>The downstream genes linked to 'GCK' are 'PKLR' and 'PK1'. We have already identified that these genes are associated with the molecular function 'pyruvate kinase activity'. Now, we need to <b>map these interactions to specific pathways and diseases to evaluate</b> their potential effects on <b>Type II diabetes mellitus</b>.</i></p>	<p>functions calling</p>	<pre>   0   properties(g2)   Link     https://www.kegg.jp/dbget-bin/www_bget?hsa:5313+hsa:5315   Name     hsa:5313 hsa:5315   Type     Gene   Aliases     &gt; 0 PKLR     &gt; 1 PK1     &gt; 2 PKL     &gt; 3 PKRL     &gt; 4 RPK...</pre>
---	--	--------------------------	---



5

**Final Answer:**

*The gene 'GCK' is associated with the disease 'Type II diabetes mellitus'. It interacts with the genes 'PKLR' and 'PK1', which are associated with the molecular function 'pyruvate kinase activity'. This suggests that these genes may play a role in the metabolic processes that are disrupted in Type II diabetes mellitus. The downstream genes linked to 'GCK' are also 'PKLR' and 'PK1'. These genes are associated with the disease 'Type II diabetes mellitus' and are part of the same pathway (path:hsa04930). Therefore, the cascade downstream effects of 'GCK' on Type II diabetes mellitus could be mediated through its interactions with 'PKLR' and 'PK1', potentially affecting the metabolic processes associated with this disease.*

## 7. System Analyze

### 7.1 Strengths

**PathwayGenie enables strategic reasoning and iterative action:** When analyzing a gene's downstream effects on diseases, the LLM iteratively generates queries, interacts with tools to gather data, and refines reasoning based on outcomes. This cyclical process ensures accurate, data-driven hypotheses, with the ReAct framework offering a structured and adaptive approach to complex biomedical challenges.

### 7.2 Future Enhancements

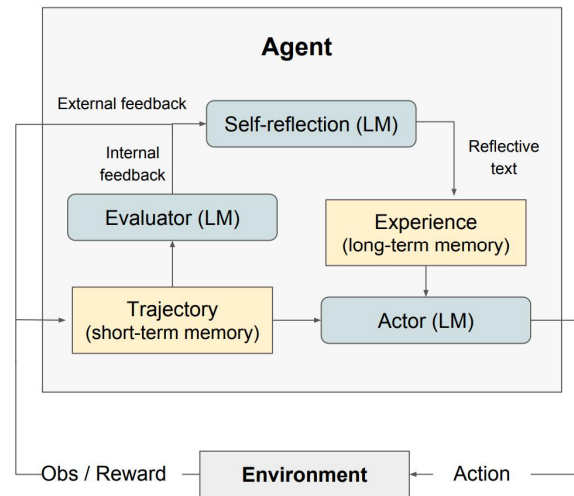
Current limitation:

- Its performance heavily relies on central model, which determines how the agent operates.
- Context window size is a limitation. Previous results are stacked in the StateManager and fed to the LLM with each iteration. For complex tasks requiring many iterations, the context window may become strained.



Based on these, the next step agent architecture improvement can be:

- Reinforcement based architecture: involving individual critic or to evaluation current situation, give informative suggestion and feedback to the central model.
- Short-term and long-term memory: similar to the way that humans remember fine-grain recent details while also recalling distilled important experiences from long-term memory.



*Picture 8 Diagram of Reflexion*

## 8. References

- [1] KEGG Markup Language (KGML) :<https://www.genome.jp/kegg/xml/docs/>
- [2] GENE Ontology Annotation File: <https://geneontology.github.io/docs/go-annotation-file-gaf-format-2.2>
- [3] ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023
- [4] LangGraph-Building language agents as graphs: <https://langchain-ai.github.io/langgraph/>
- [5] Langsmith: <https://www.langchain.com/langsmith>
- [6] Reflexion: Language Agents with Verbal Reinforcement Learning. NeurIPS 2024