



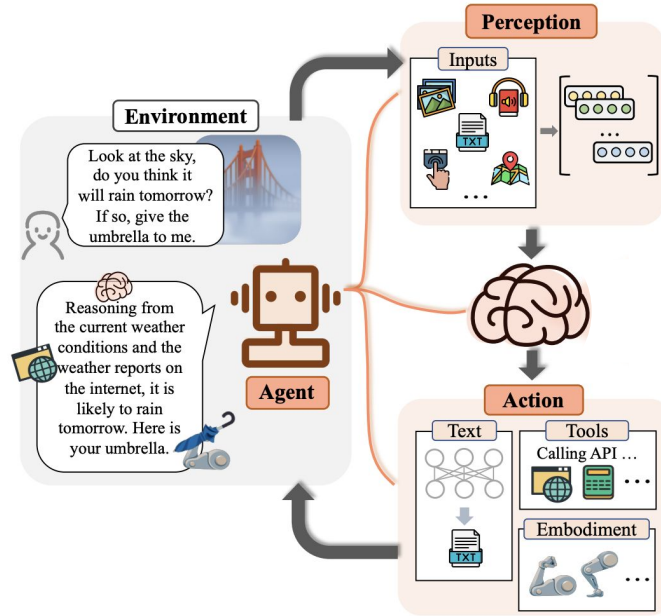
# AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Awadallah, Ryen W. White, Doug Burger, Chi Wang

# What are Agents?



An “*Agent*” is an entity with the capacity to act in a given environment for task solving.



## **Conversable Agents**

A conversable agent is an entity with a specific role that can pass messages to send and receive information to and from other conversable agents, e.g., to start or continue a conversation.

## **Conversable Agents**

A conversable agent is an entity with a specific role that can pass messages to send and receive information to and from other conversable agents, e.g., to start or continue a conversation.

## **Why can LLMs be used as Agents?**

## Conversable Agents

A conversable agent is an entity with a specific role that can pass messages to send and receive information to and from other conversable agents, e.g., to start or continue a conversation.

## Why can LLMs be used as Agents?

**Autonomy:** Autonomy in agents refers to their ability to operate independently without human intervention, including following instructions, initiating actions, and adapting outputs based on environmental input. LLMs demonstrate autonomy by generating human-like text, engaging in conversations, performing tasks, and showcasing creativity.

**Reactivity:** Reactivity in agents refers to their ability to quickly respond to changes in the environment by perceiving and taking appropriate actions. Language models traditionally have limited perceptual and action spaces but can expand them using tool usage, i.e, querying APIs to perform tasks. LLM based agents can exhibit reasoning and planning abilities comparable to symbolic agents through techniques like Chain-of-Thought (CoT) and problem decomposition

# Building LLM Agents - ReAct



## (1) Hotspot QA

**Question:** Aside from the Apple Remote, what other device can control the program Apple Remote was originally designed to interact with?

### (1a) Standard

**Answer:** iPod



### (1b) CoT (Reason Only)

**Thought:** Let's think step by step. Apple Remote was originally designed to interact with Apple TV. Apple TV can be controlled by iPhone, iPad, and iPod Touch. So the answer is iPhone, iPad, and iPod Touch.

**Answer:** iPhone, iPad, iPod Touch



### (1c) Act-Only

**Act 1:** Search[Apple Remote]  
**Obs 1:** The Apple Remote is a remote control ...

**Act 2:** Search[Front Row]  
**Obs 2:** Could not find [Front Row]. Similar: ...

**Act 3:** Search[Front Row (software)]  
**Obs 3:** Front Row is a discontinued media center software ...

**Act 4:** Finish[yes]



## (1d) ReAct (Reason + Act)

**Thought 1:** I need to search Apple Remote and find the program it was originally designed to interact with.

**Act 1:** Search[Apple Remote]

**Obs 1:** The Apple Remote is a remote control introduced in October 2005 by Apple ... originally designed to control the Front Row media center program ...

**Thought 2:** Apple Remote was originally designed to control the Front Row media center program. I need to search Front Row next and find what other device can control it.

**Act 2:** Search[Front Row]

**Obs 2:** Could not find [Front Row]. Similar: ['Front Row Seat to Earth', 'Front Row Motorsports', 'Front Row (software)', ...]

**Thought 3:** Front Row is not found. I need to search Front Row (software) .

**Act 3:** Search[Front Row (software) ]

**Obs 3:** Front Row is a discontinued media center software ...

**Thought 4:** Front Row (software) is controlled by an Apple Remote or the keyboard function keys. So the answer is keyboard function keys.

**Act 4:** Finish[keyboard function keys]

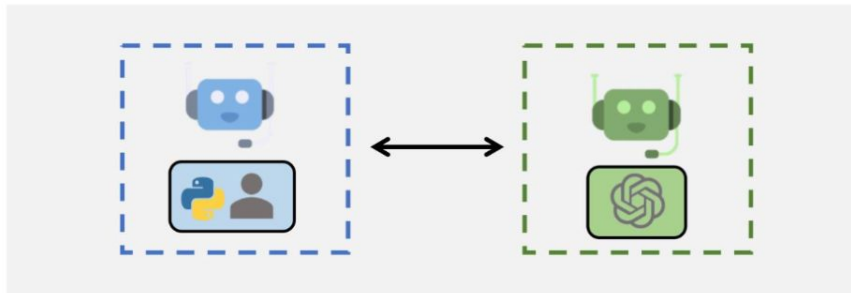


# AutoGen



How to empower every developer to build conversable agents?

*A programming framework for easily building multi-agent applications that could span a broad spectrum of domains and complexities*

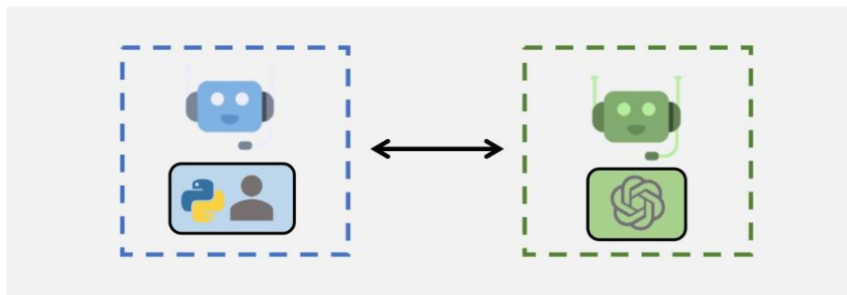


- Initially developed in FLAML (Nov 2022)
- Spined off to a standalone repo (October 2023)

[aka.ms/autogen](https://aka.ms/autogen)

How to empower every developer to build conversable agents?

## Hugging Face 🤗 for Agents



- Initially developed in FLAML (Nov 2022)
- Spined off to a standalone repo (October 2023)

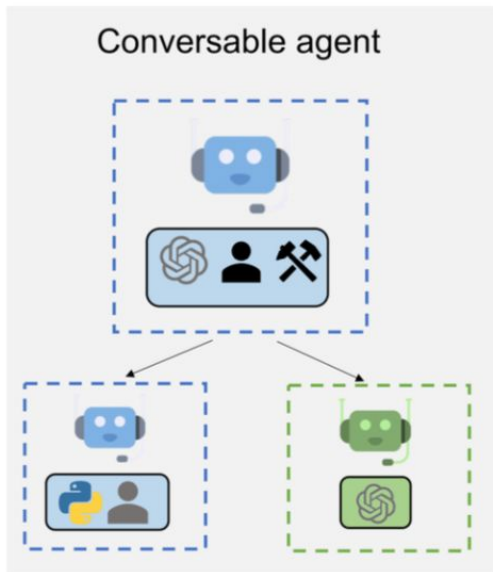
[aka.ms/autogen](https://aka.ms/autogen)



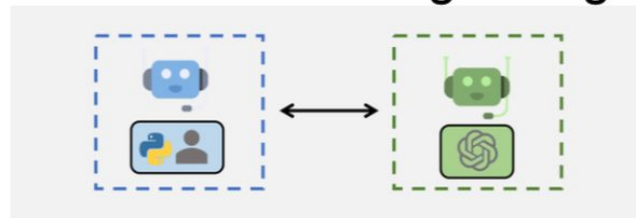
**Aim:** Multi-agent conversation framework with generic abstraction and effective implementation that has the flexibility to satisfy different application needs.

- How can we design individual agents that are capable, reusable, customizable, and effective in multi-agent collaboration?
- How can we develop a straightforward, unified interface that can accommodate a wide range of agent conversation patterns?

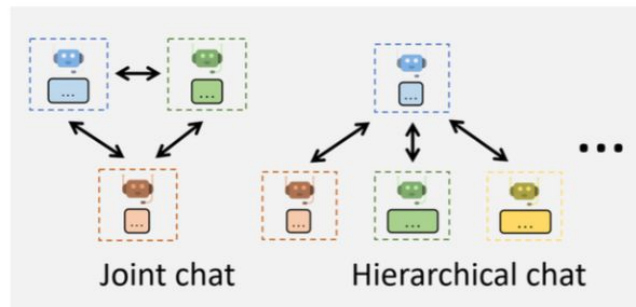
Define agents:  
Conversable & Customizable



Get them to talk:  
Conversation Programming



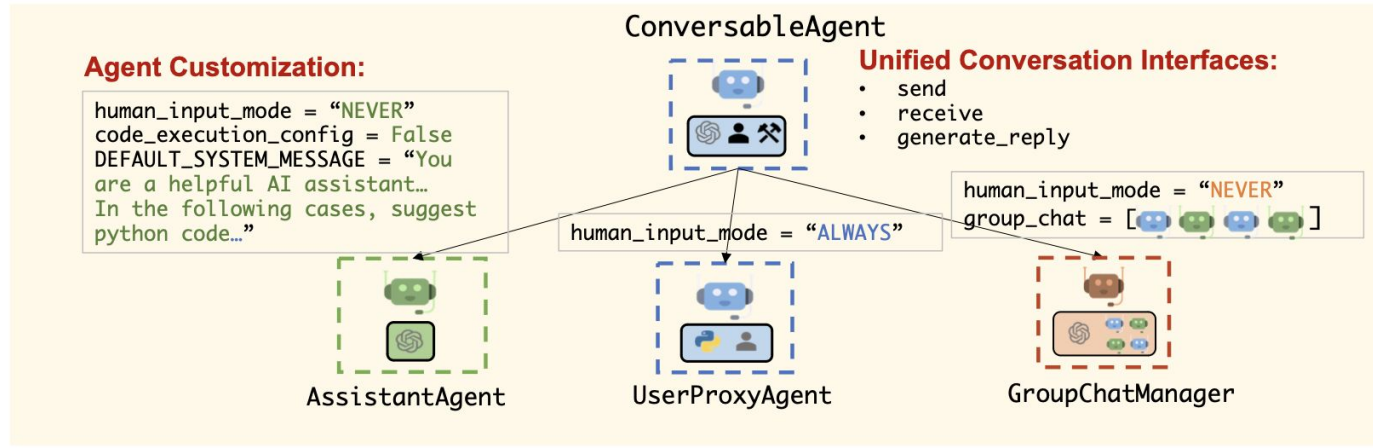
Multi-Agent Conversations



Flexible Conversation Patterns

A conversable agent is an entity with a specific role that can pass messages to send and receive information (in the form of text) to and from other agents, e.g., to start or continue a conversation

A conversable agent is an entity with a specific role that can pass messages to send and receive information (in the form of text) to and from other agents, e.g., to start or continue a conversation





Involves defining agent capabilities and roles, and programming their interaction behaviors through conversation-centric computation and control.


**Computation** refers to the actions agents undertake to compute their responses within a multi-agent conversation, while control flow dictates the sequence or conditions under which these computations occur. AutoGen's approach focuses on conversation-centric computations, where agents engage in actions pertinent to their conversations, leading to message passing for subsequent interactions unless certain termination conditions are met.

**Control flow** in AutoGen is driven by conversations, meaning that agents' decisions on message recipients and computation procedures are influenced by the ongoing inter-agent dialogue. This paradigm facilitates intuitive reasoning about complex workflows through agent actions and conversation-driven message passing.



# Conversation Programming




Plot a chart of META and TESLA stock price change YTD.





Execute the following code...





Error package yfinance is not installed




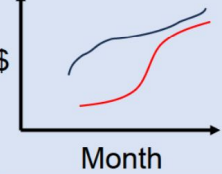
Sorry! Please first pip install yfinance and then execute the code





Installing...




Output:





...




No, please plot % change!



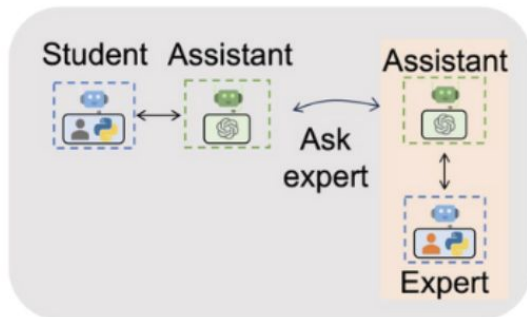
Here is the revised code ...



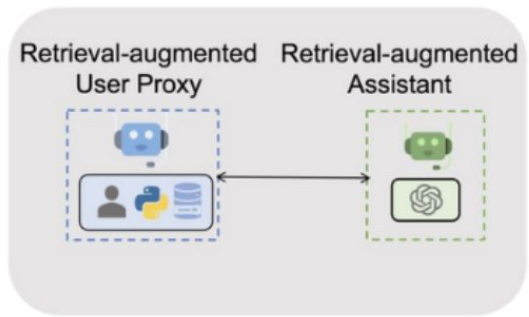
Output:



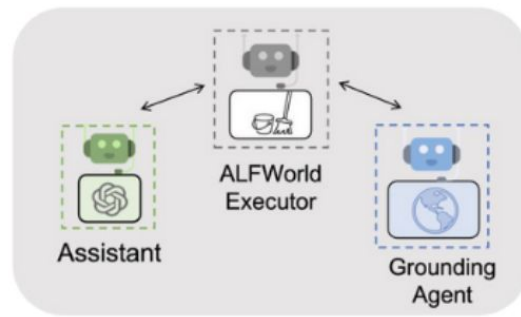
# Application Enabled by AutoGens



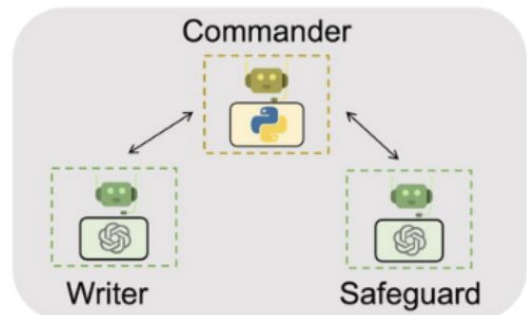
A1. Math Problem Solving



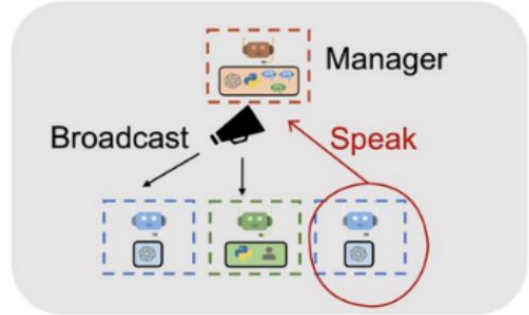
A2. Retrieval-augmented Chat



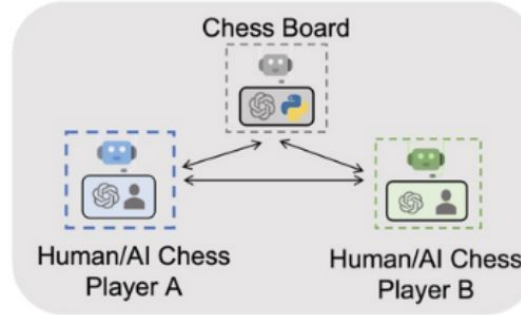
A3. Decision Making



A4. Multi-agent Coding



A5. Dynamic Group Chat



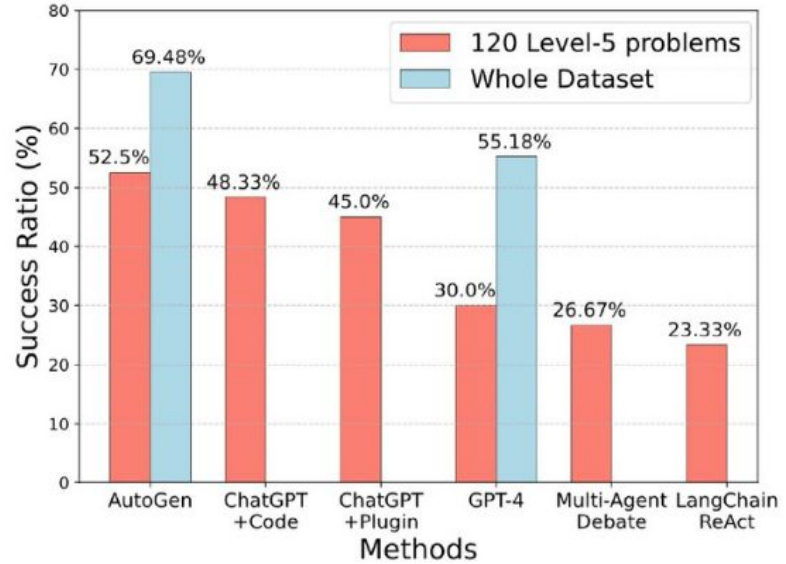
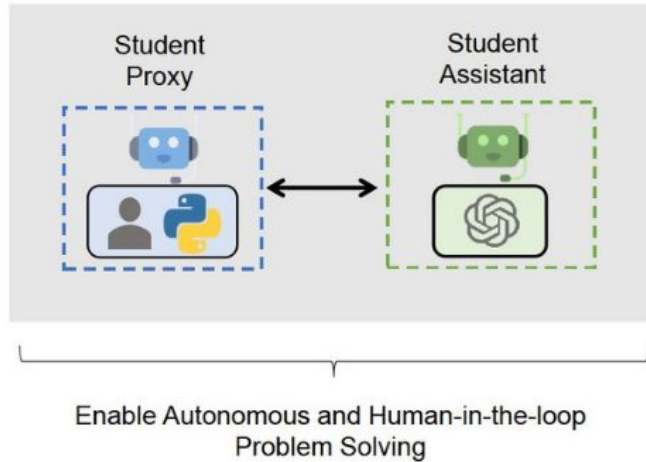
A6. Conversational Chess



# Math Problem Solving



# Math Problem Solving

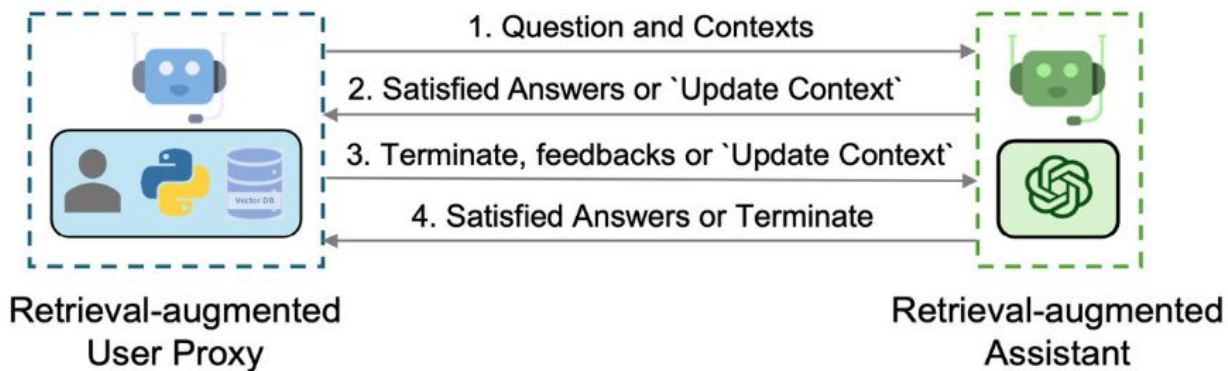


Autonomous Problem-Solving Performance



# Retrieval Augmented Chat & Question Answering

# Retrieval-Augmented Chat



- Retrieval Augmentation
- Interactive Retrieval

### Task description

who carried the usa flag in opening ceremony?

#### (a) W/O Interactive Retrieval

**UserProxy:** You're a retrieve augmented chatbot. You answer user's questions based on your own knowledge and the context provided by the user. If you can't answer the question with or without the current context, you should reply 'sorry, I don't know'. You must give as short an answer as possible.

User's question is: who carried the usa flag in opening ceremony  
Context is: < P > On January 17 , 1899 , under ... The Astros beat the Yankees that night , 2 - 1 . < /P >

**Assistant:** Sorry, I cannot find any information about who carried the USA flag in the opening ceremony. Sorry, I don't know.



#### (b) W/ Interactive Retrieval

**UserProxy:** You're a retrieve augmented chatbot. You answer user's questions based on your own knowledge and the context provided by the user. If you can't answer the question with or without the current context, you should reply exactly 'UPDATE CONTEXT'. You must give as short an answer as possible.

User's question is: who carried the usa flag in opening ceremony  
Context is: < P > On January 17 , 1899 , under ... The Astros beat the Yankees that night , 2 - 1 . < /P >

**Assistant:** Sorry, I cannot find any information about who carried the USA flag in the opening ceremony. UPDATE CONTEXT.

**UserProxy:** ...User's question is: who carried the usa flag in opening ceremony

Context is: < T able > ... Erin Hamlin ... < /T d >

**Assistant:** Erin Hamlin carried the USA flag in the opening ceremony.



Retrieval-augmented Chat without vs. with *interactive retrieval*

## Task description

who carried the usa flag in opening ceremony?

### (a) W/O Interactive Retrieval

**UserProxy:** You're a retrieve augmented chatbot. You answer user's questions based on your own knowledge and the context provided by the user. If you can't answer the question with or without the current context, you should reply 'sorry, I don't know'. You must give as short an answer as possible.

User's question is: who carried the usa flag in opening ceremony  
Context is: < P > On January 17 , 1899 , under ... The Astros beat the Yankees that night , 2 - 1 . < /P >

**Assistant:** Sorry, I cannot find any information about who carried the USA flag in the opening ceremony. Sorry, I don't know.



### (b) W/ Interactive Retrieval

**UserProxy:** You're a retrieve augmented chatbot. You answer user's questions based on your own knowledge and the context provided by the user. If you can't answer the question with or without the current context, you should reply exactly 'UPDATE CONTEXT'. You must give as short an answer as possible.

User's question is: who carried the usa flag in opening ceremony  
Context is: < P > On January 17 , 1899 , under ... The Astros beat the Yankees that night , 2 - 1 . < /P >

**Assistant:** Sorry, I cannot find any information about who carried the USA flag in the opening ceremony. UPDATE CONTEXT.

**UserProxy:** ...User's question is: who carried the usa flag in opening ceremony

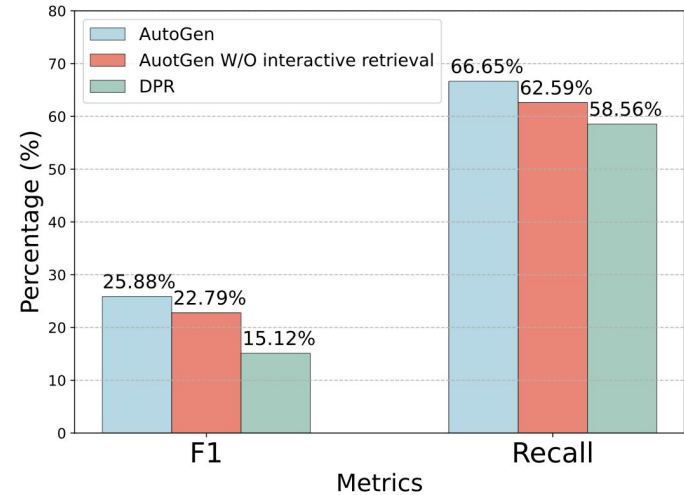
Context is: < T able > ... Erin Hamlin ... < /T d >

**Assistant:** Erin Hamlin carried the USA flag in the opening ceremony.



## Retrieval-augmented Chat without vs. with *interactive*

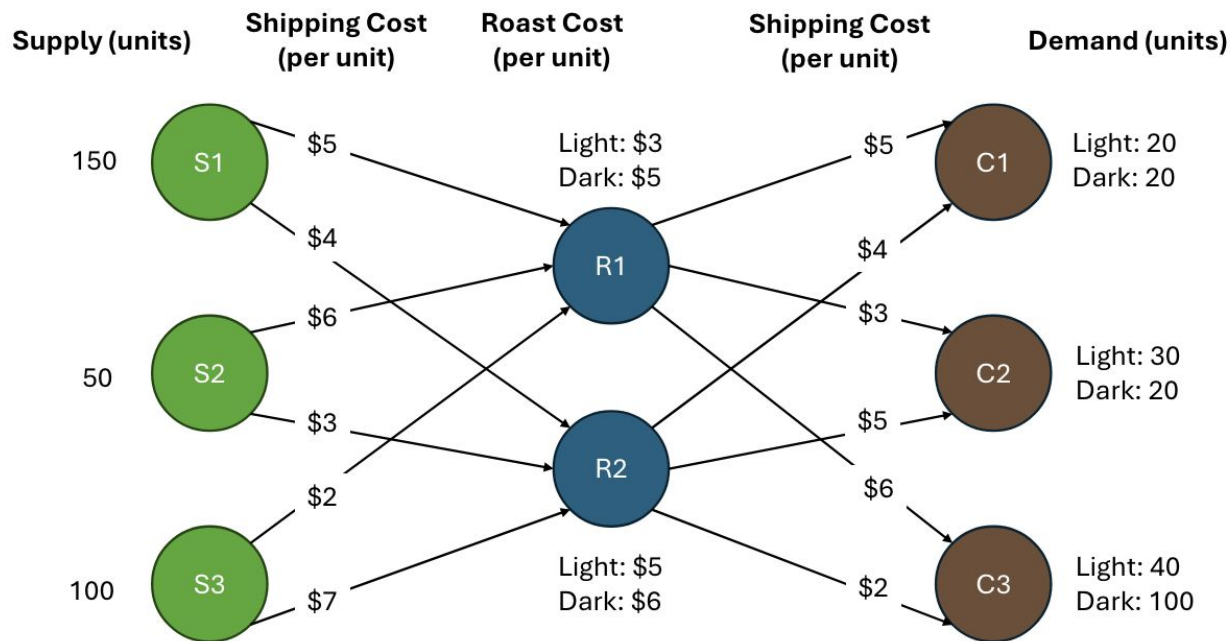
AutoGen introduces an interactive retrieval feature where, if the retrieved context lacks information, the assistant responds with "Sorry, I cannot find any information about... UPDATE CONTEXT," prompting further retrieval attempts





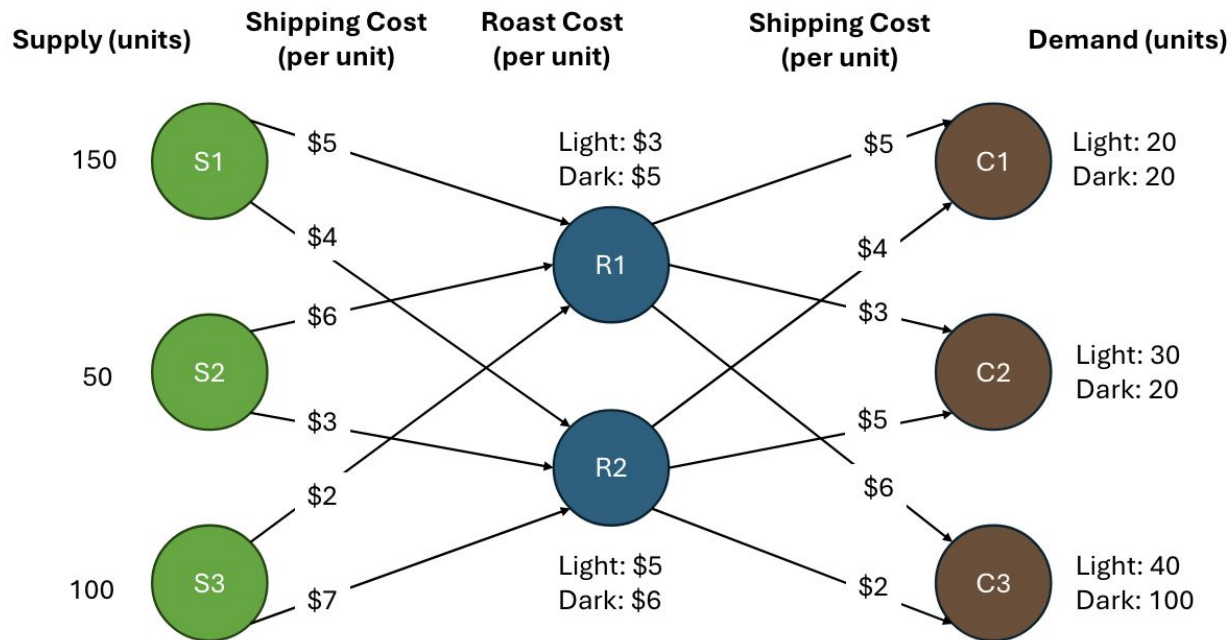
# Multi-Agent Coding

# Task - Optimization and Planning



(a) Problem setup.

# Task - Optimization and Planning



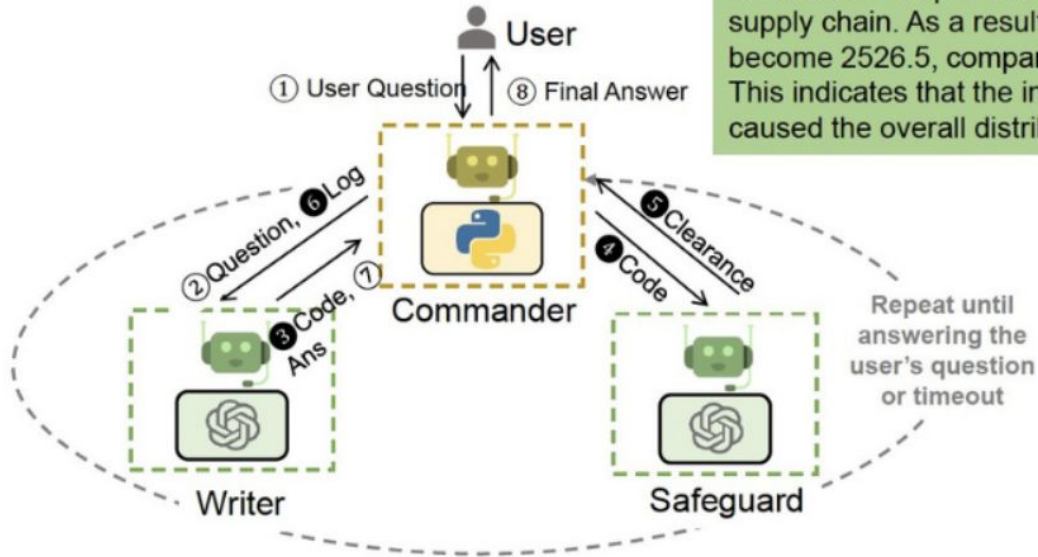
(a) Problem setup.

What if the roasting cost is increased by 5% because of the potential salary increase ?



What if the roasting cost is increased by 5% because of the potential salary increase?

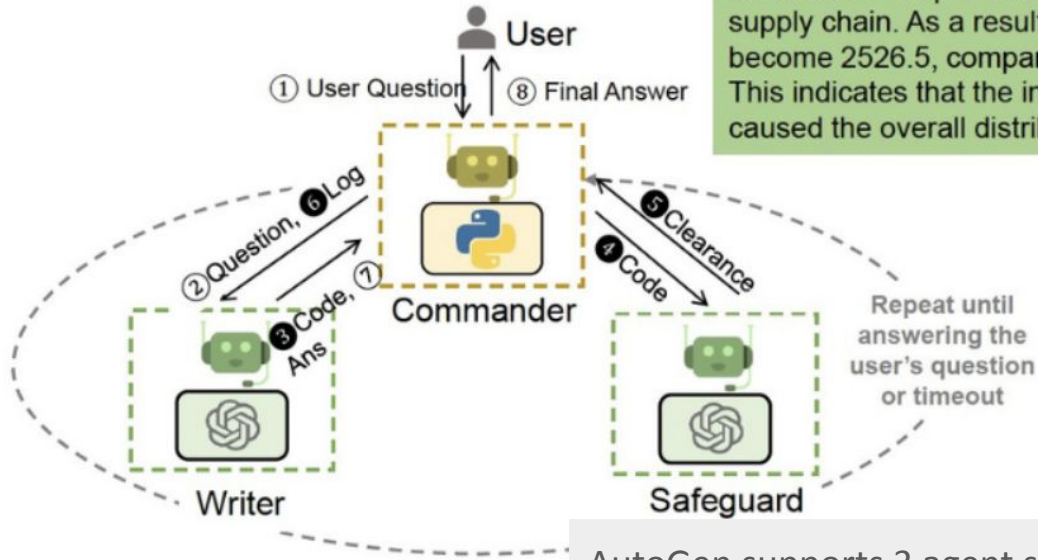
After increasing the roasting costs by 5%, I've re-calculated the optimal coffee distribution solution for the supply chain. As a result, the new optimal total cost has become 2526.5, compared to the initial cost of 2470.0. This indicates that the increased roasting costs have caused the overall distribution cost to rise by 56.5 units.



Writer, combines the functions of a “Coder” and an “Interpreter”

What if the roasting cost is increased by 5% because of the potential salary increase?

After increasing the roasting costs by 5%, I've re-calculated the optimal coffee distribution solution for the supply chain. As a result, the new optimal total cost has become 2526.5, compared to the initial cost of 2470.0. This indicates that the increased roasting costs have caused the overall distribution cost to rise by 56.5 units.

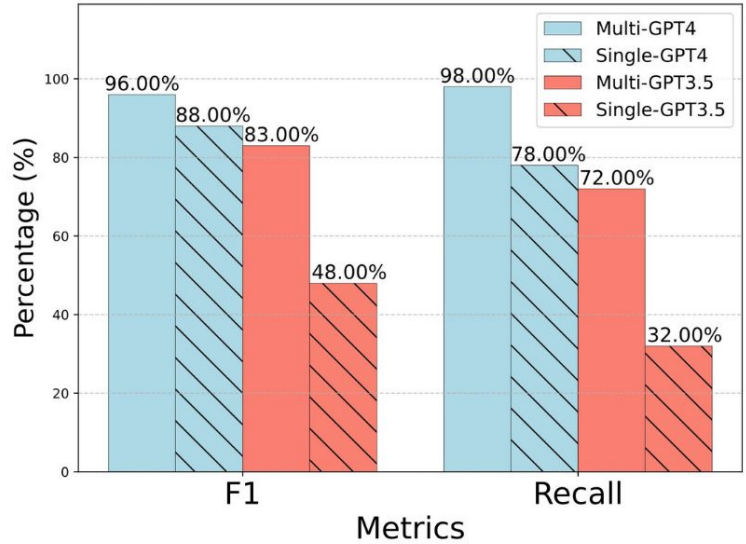


Writer, combines the functions of a “Coder” and an “Interpreter”

AutoGen supports 3 agent setups due to its customizable framework, additionally also allows dynamic switching between code and text

What if the roasting cost is increased by 5% because of the potential salary inc

•Safeguard performance



ing costs by 5%, I've re-  
ree distribution solution for the  
the new optimal total cost has  
d to the initial cost of 2470.0.  
reased roasting costs have  
ition cost to rise by 56.5 units.

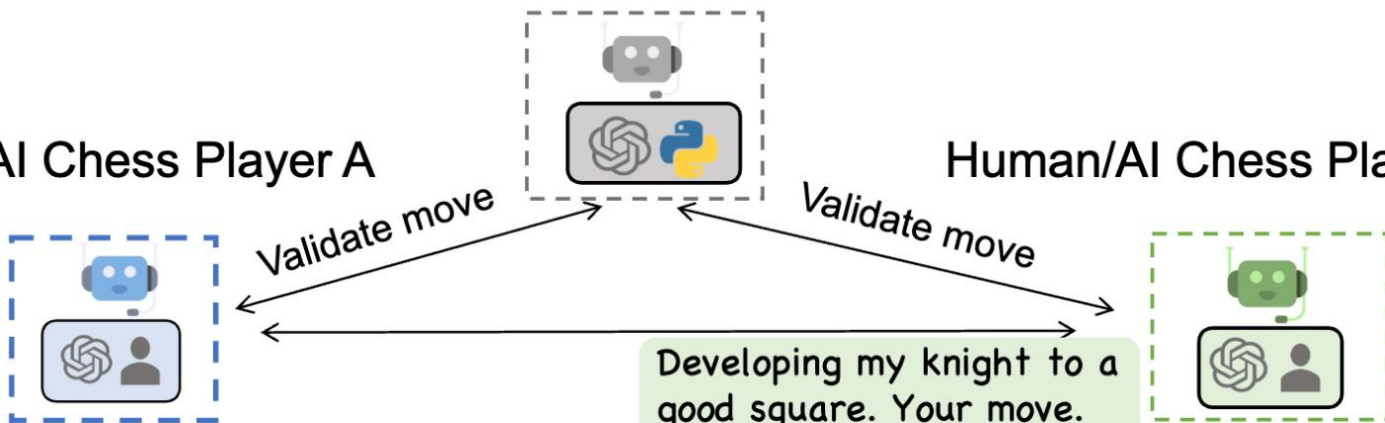


# Conversational Chess

# Chess Board

Human/AI Chess Player A

Human/AI Chess Player B



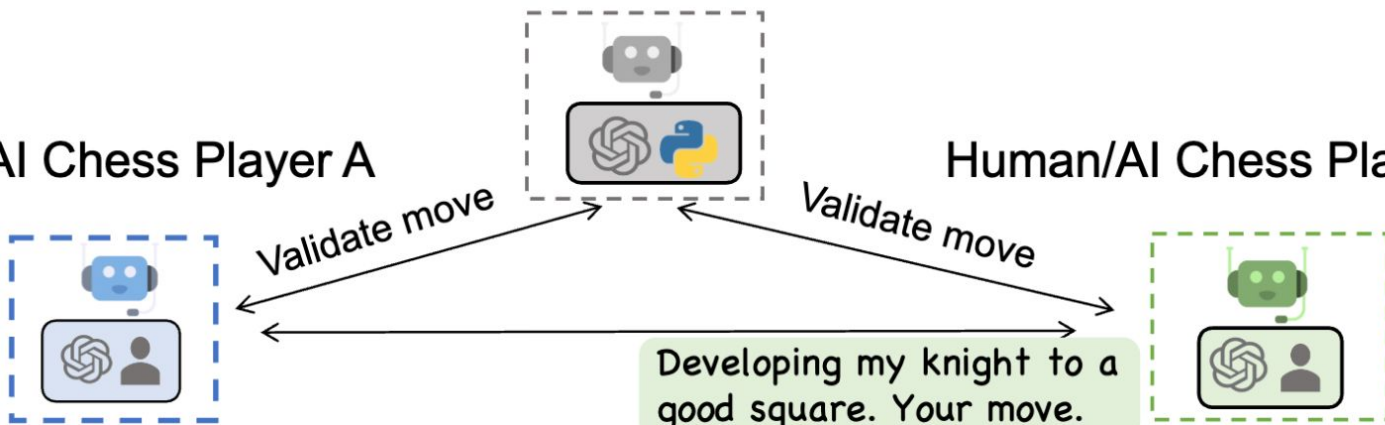
Challenging your pawn in the center. Your move.



# Chess Board

Human/AI Chess Player A

Human/AI Chess Player B



Challenging your pawn in the center. Your move.

Developing my knight to a good square. Your move.



register\_reply method supported by AutoGen agents to instantiate player agents and a board agent with custom reply functions., so players can only make moves while the chess board agent is restricted to only verify the legality of the move.

(a) W/O Board Agent

Player White: Your move  
 Player black: Alright, let's start. I'll move my pawn from e7 to e5.  
 Here's how the board looks now:  
 r n b q k b n r  
 p p p p . p p p  
 . . . . .  
 . . . . p . . . .  
 . . . . .  
 P P P P P P P P  
 R N B Q K B N R  
 Your turn!

...

Player White: Yes, it certainly is! Let's keep this up.  
 I'll move my knight from b8 to c6. Here's the updated board:  
 r n b q k b . r  
 P P P P . p p p  
 . . . . n . . . .  
 . . . . p . . . .  
 . . . . P . . . .  
 . . . . N . . . .  
 P P P P . P P P  
 R N B Q K B . R  
 Your move! You're giving me a tough time here.



(b) W/ Board Agent

Player White: Your move  
 Player black: Alright, let's start. I'll move my pawn from e7 to e5.  
 Here's how the board looks now:  
 r n b q k b n r  
 P P P P . P P P  
 . . . . .  
 . . . . p . . . .  
 . . . . .  
 P P P P P P P P  
 R N B Q K B N R  
 Your turn!

Player White: Yes, it certainly is! Let's keep this up. I'll move my knight from b8 to c6. Here's the updated board:  
 r n b q k b . r  
 P P P P . P P P  
 . . . . n . . . .  
 . . . . P . . . .  
 . . . . P . . . .  
 . . . . N . . . .  
 P P P P . P P P  
 R N B Q K B . R  
 Your move! You're giving me a tough time here.  
 Board Agent: Your move is illegal. You changed the rock at a8 to knight and move it to c6. Please check your decision and re-make your move.  
 Player White: Sorry for the mistakes. I'll move my knight from b8 to c6.  
 Here's the updated board:

r . b q k b . r  
 P P P P . P P P  
 . . n . . n . . .  
 . . . . P . . . .  
 . . . . P . . . .  
 . . . . N . . . .  
 P P P P . P P P  
 R N B Q K B . R



## Summary of Benefits

---

Multi-agent  
conversation

Improve over natural iteration

---

Divide & conquer

---

Grounding & validation

---

AutoGen

Modularity for ease of use & extension

---

Natural human participation

---

Fast & creative experimentation

---

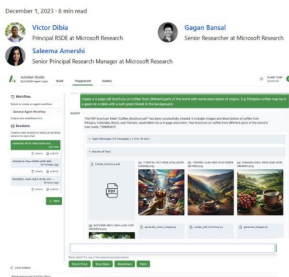


- **Finite context length:** The restricted context capacity limits the inclusion of historical information, detailed instructions, API call context, and responses. The design of the system has to work with this limited communication bandwidth, while mechanisms like self-reflection to learn from past mistakes would benefit a lot from long or infinite context windows. Although vector stores and retrieval can provide access to a larger knowledge pool, their representation power is not as powerful as full attention.
- **Challenges in long-term planning and task decomposition:** Planning over a lengthy history and effectively exploring the solution space remain challenging. LLMs struggle to adjust plans when faced with unexpected errors, making them less robust compared to humans who learn from trial and error. How many previous steps do we unroll ?

- How to design optimal multi-agent workflows?
  - Evaluation
  - Optimization
- How to create highly capable agents?
  - Integration
  - Learning
- How to enable scale, safety and human agency?
  - Teaching
  - Interface

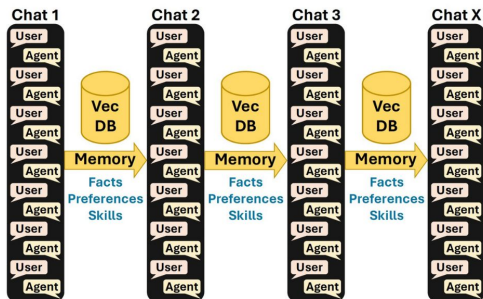
## AutoGen Studio: Interactively Explore Multi-Agent Workflows

- Declaratively define and modify agents and multi-agent workflows
- Create chat sessions with the specified agents and view results
- Explicitly add skills to your agents and accomplish more tasks
- Publish your sessions to a local gallery



## MultiModal

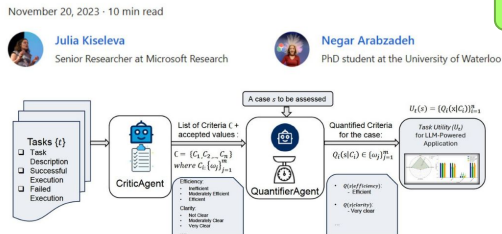
## Teachable Agents remember user teachings long-term



## Longer contexts

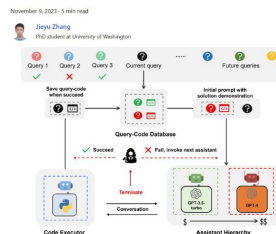
## AgentEval - Assess the Utility of LLM-powered Applications

## Evaluation



## EcoAssistant - Using LLM Assistants More Accurately and Affordably

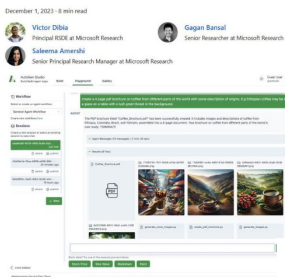
issistant is designed to solve user is more accurately and lably  
 ize the cost of using GPT models istant Hierarchy  
 ritage the idea of Retrieval- ent Generation (RAG) to roved the success rate viaSolution nstration



## Learning Agents

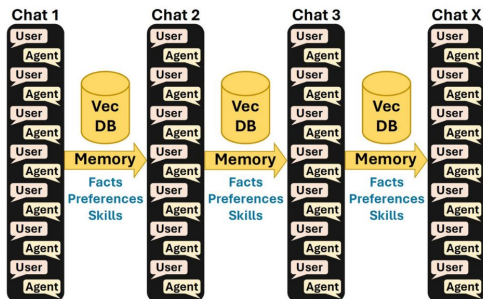
## AutoGen Studio: Interactively Explore Multi-Agent Workflows

Declaratively define and modify agents and multi-agent workflows  
 Create chat sessions with the specified agents and view results  
 Explicitly add skills to your agents and accomplish more tasks  
 Publish your sessions to a local gallery



MultiModal

## Teachable Agents remember user teachings long-term



Longer contexts

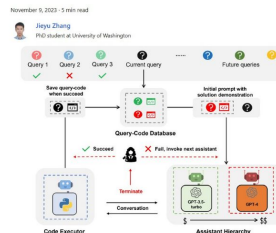
## AgentEval - Assess the Utility of LLM-powered Applications



Evaluation

## EcoAssistant - Using LLM Assistants More Accurately and Affordably

Assistant is designed to solve user is more accurately and lably  
 ize the cost of using GPT models  
 istant Hierarchy  
 range the idea of Retrieval-ent Generation (RAG) to  
 prove the success rate via Solution  
 nstration



Learning Agents

Further works can investigate which strategies, such as agent topology and conversation patterns, lead to the most effective multi-agent conversations





**Thank You!!**