Integrating Multi-agent dynamic and Tool-enhanced reasoning in LLMs

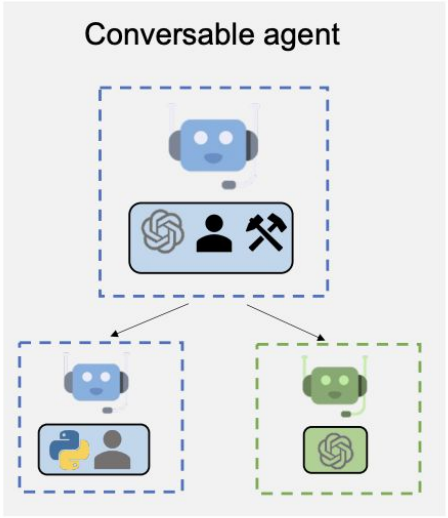
Jiayu Huang
03/05/2024

Introduction

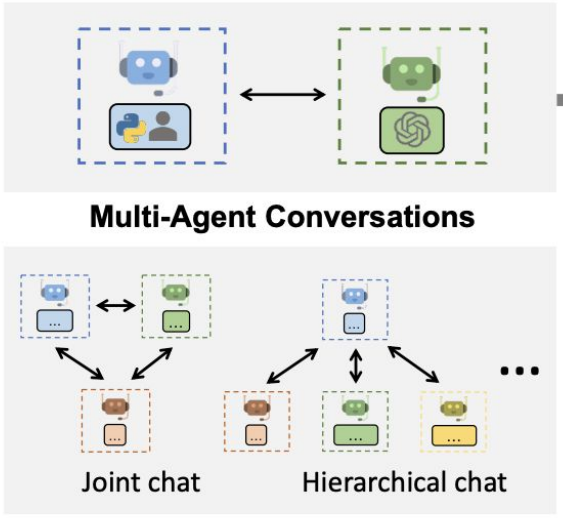
- Current LLMs struggle with complex reasoning, multi-step problem-solving, and integrating external knowledge or computations.
- Innovative frameworks: **AutoGen** facilitates multi-agent conversations to enhance LLM application, while **ART** focused on automatic multi-step reasoning and integrating tool use.
- Pros over RAG and CoT: Enhancing complex reasoning, offering improved adaptability and flexibility, and reducing human intervention.

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

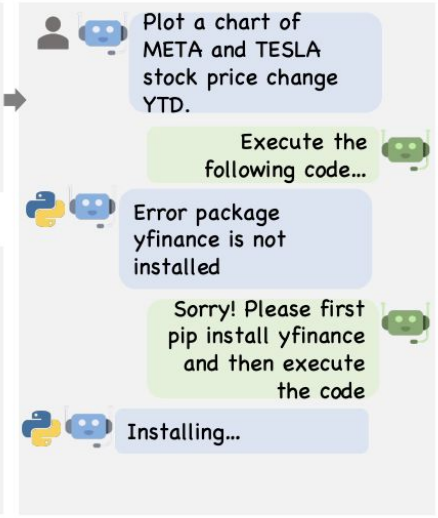
- Wu et al.,. (2023)
- Customizable and conversable (receive, react, response) agents
- Conversation programming

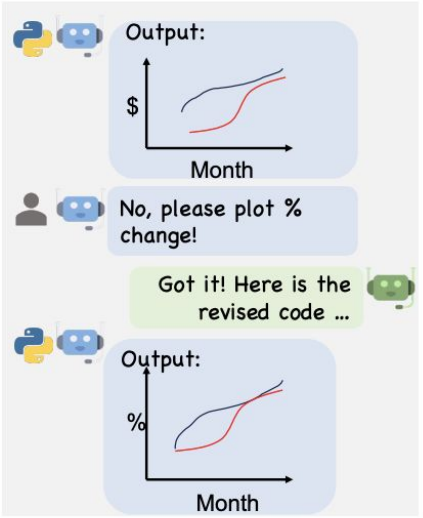


Agent Customization



Flexible Conversation Patterns





Example Agent Chat

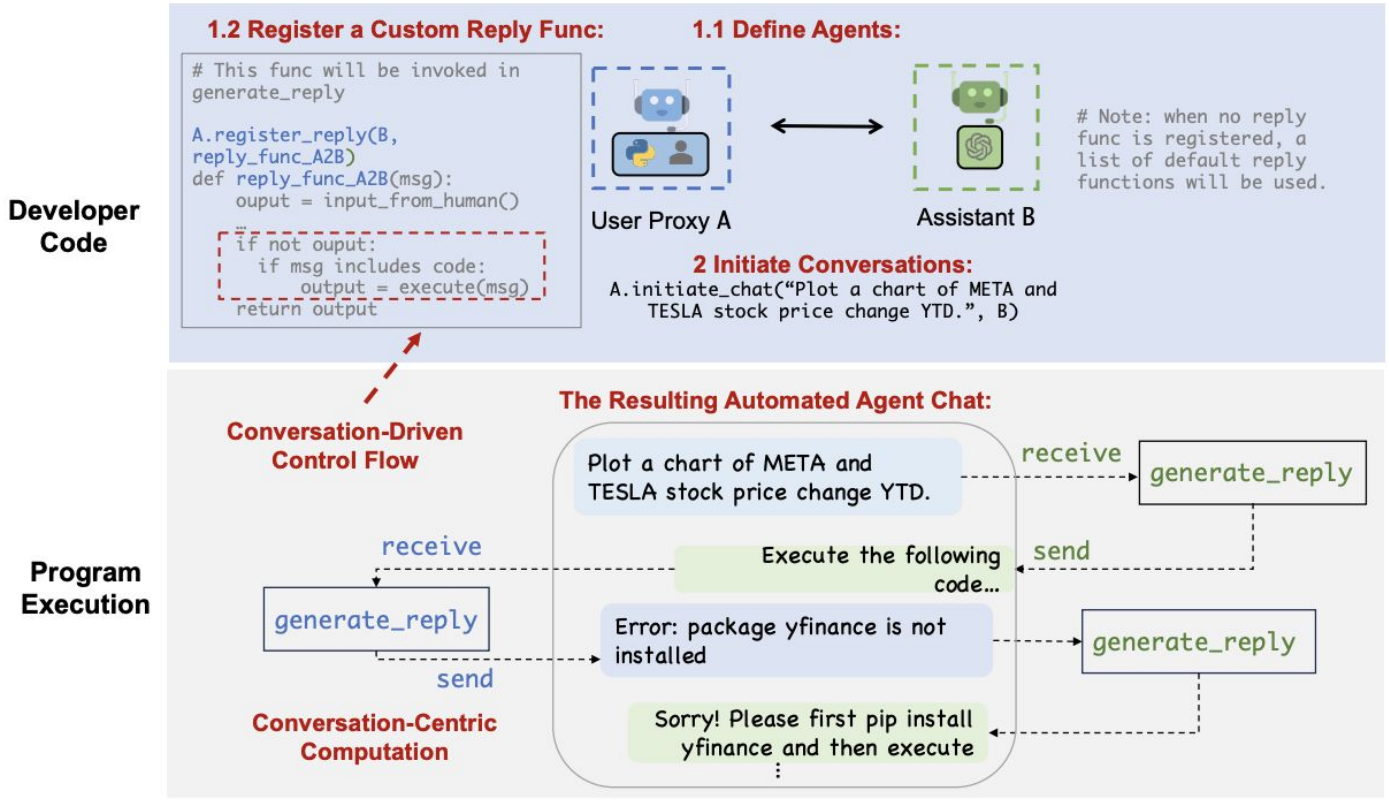
Conversation Programming

- **Computation**: Agents compute responses based on their role.
- Conversation-centric actions: Agents' actions are relevant to their conversations, facilitating message passing for further interactions.
- **Control Flow**: Sequence or conditions under which computation occur.
- Conversation-driven control flow: Agents decide whom to send messages based on the ongoing conversation.

Conversation Programming

- Unified interfaces & Auto-reply mechanisms
  - Simplified agent interaction through send/receive and reply generation functions.
  - Auto-reply for continuous conversation flow, customizable with reply functions.
- Control Fusion
  - Programming and natural language for managing control flow.
  - Flexible transitions between code and natural language controls.

Conversation Programming



Applications of AutoGen

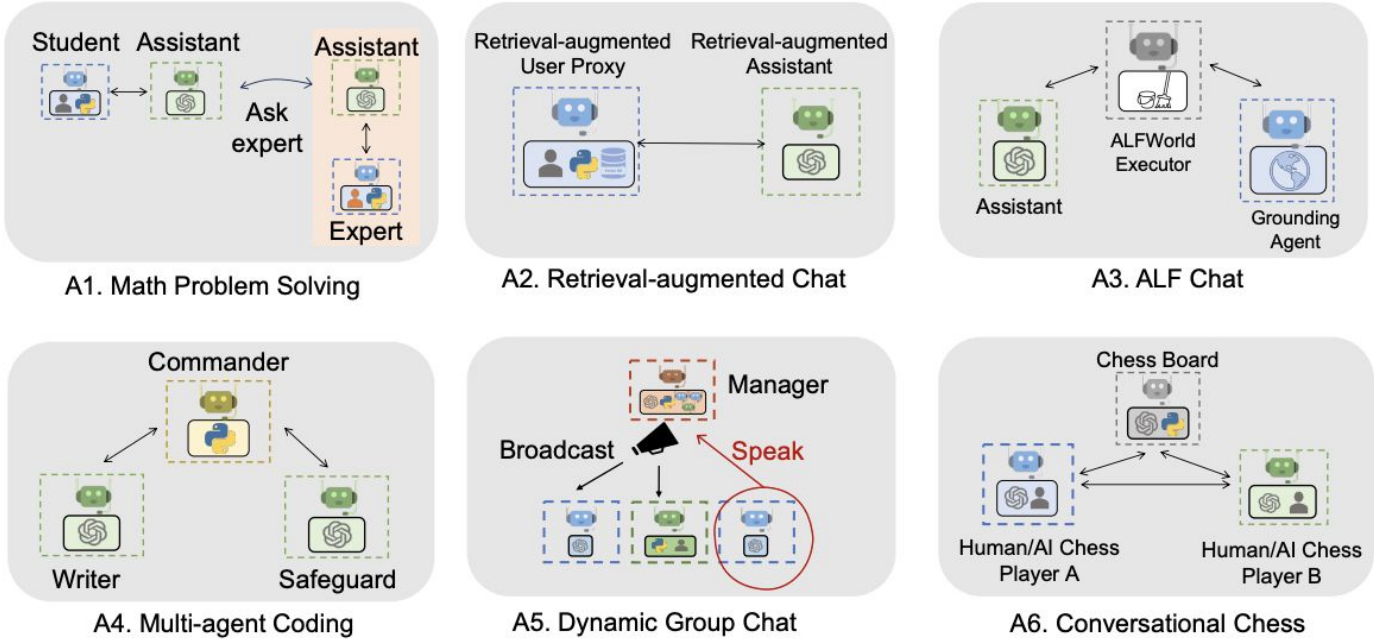
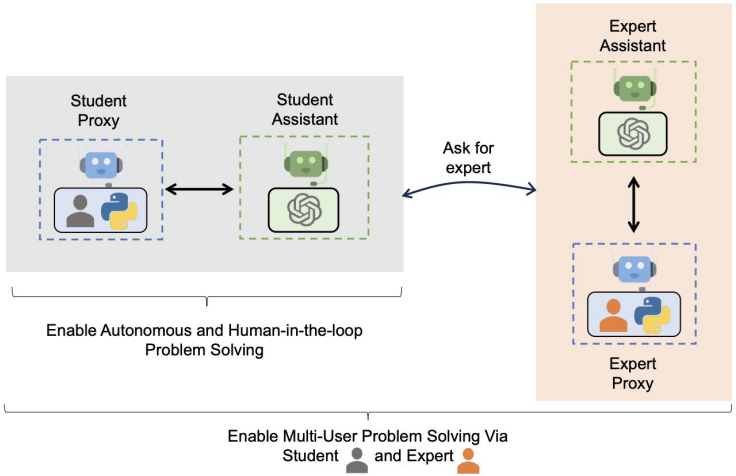
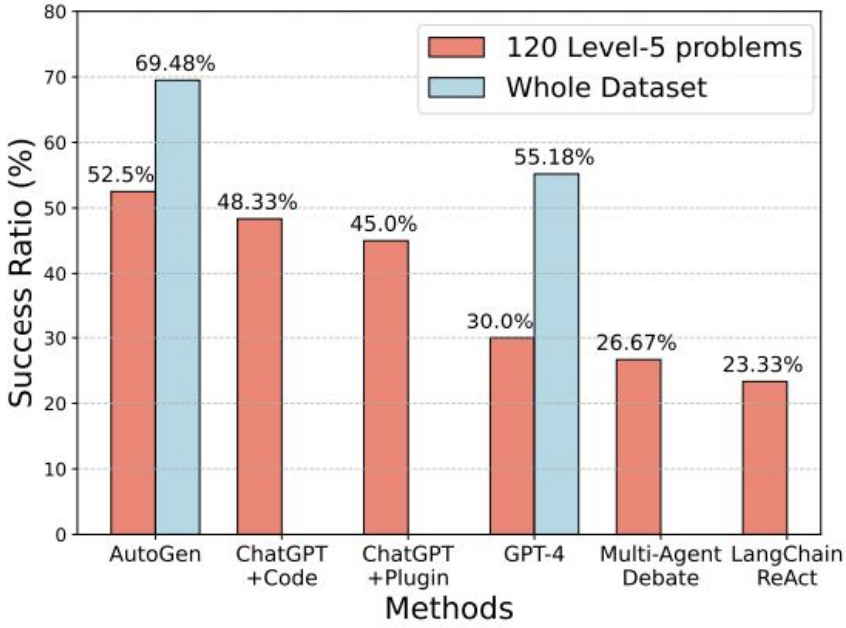


Figure 3: Six examples of diverse applications built using `AutoGen`. Their conversation patterns show `AutoGen`'s flexibility and power.

Math Problem Solving

- A system for autonomous math problem solving by directly reusing two built-in agents from AutoGen.

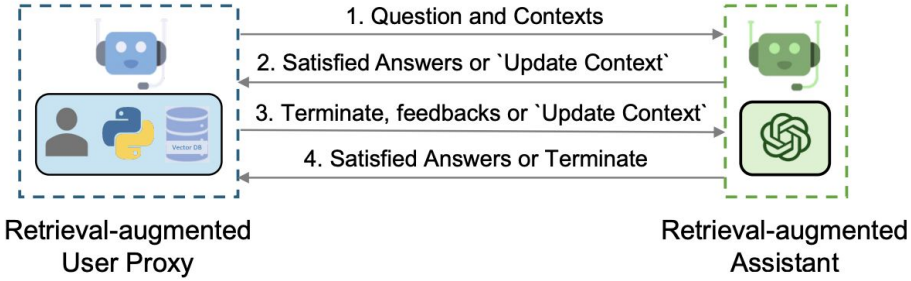


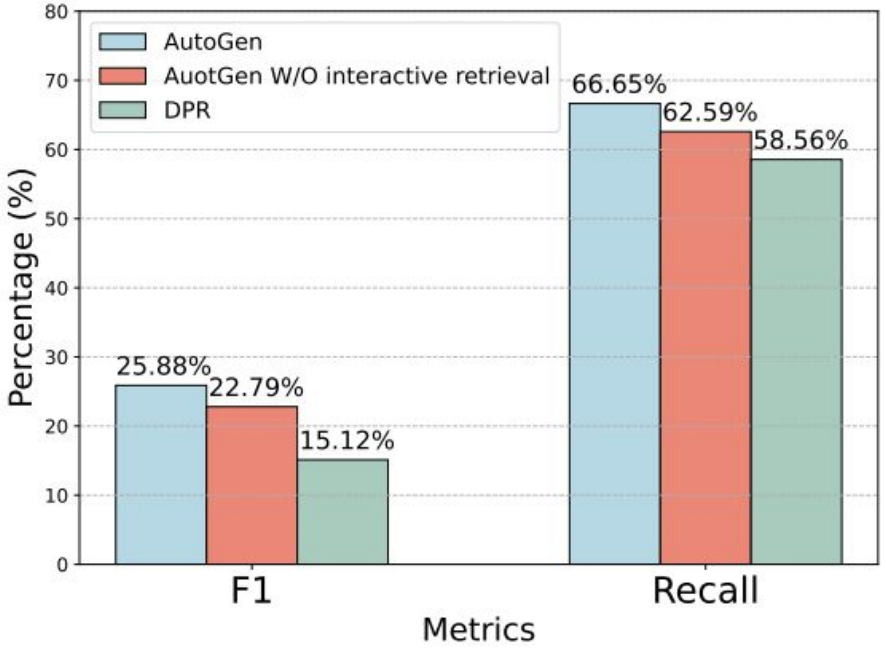


(a) A1: Performance on MATH (w/ GPT-4).

Retrieval-Augmented Question Answering

- Assess natural QA using the Natural Questions dataset, highlighting AutoGen's novel interactive retrieval feature that enhances retrieval attempts.
- “UPDATE CONTEXT”

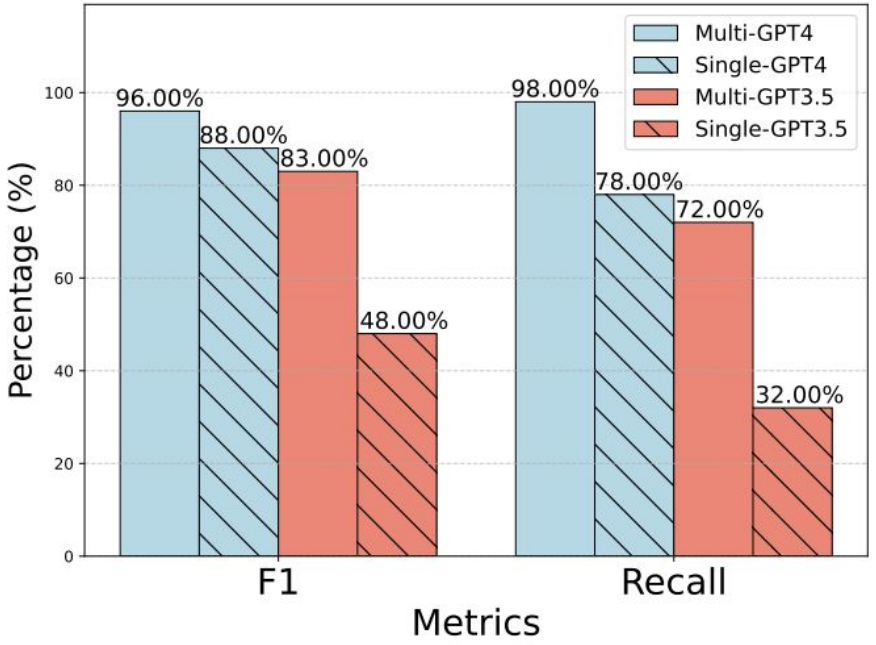




(b) A2: Q&A tasks (w/ GPT-3.5).

Multi-agent Coding

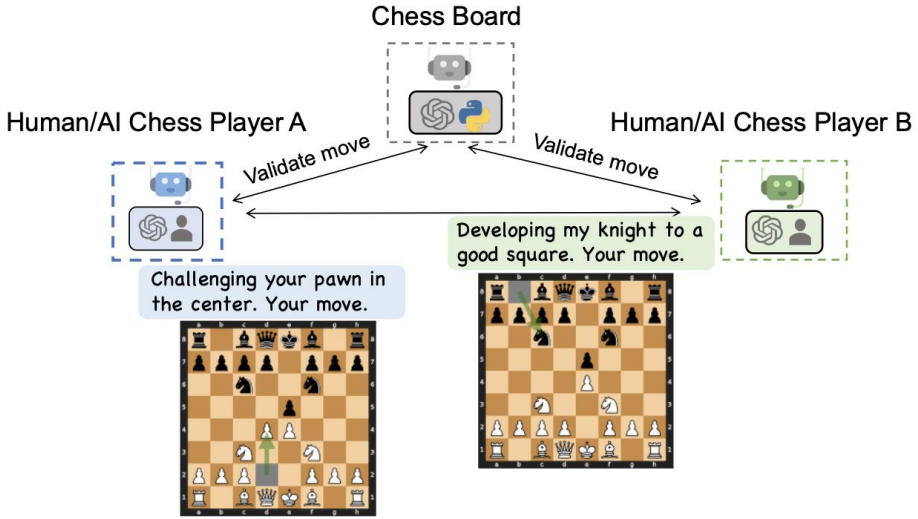
- AutoGen reduces workflow code of OptiGuide from 430 to 100 lines and streamlines optimization solution interpretation.
- Multi-agent approach increases safety and efficiency, saving users 3x time and reducing interactions by 3-5 times.



(d) A4: Performance on OptiGuide.

Conversational Chess

- Emphasizes natural, engaging gameplay through customizable agent, enabling seamless mode switching and maintaining game integrity by validating each move's legality.
- Removing the board agent will negatively affect gameplay and preventing illegal moves.



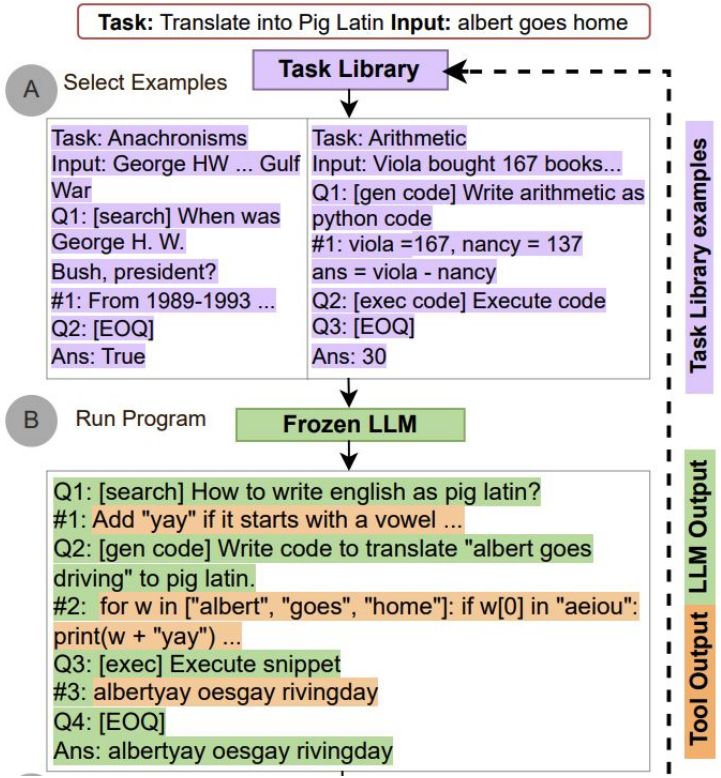
Future directions for AutoGen

- Integration and Customization: Enhance AutoGen to integrate and customize existing agents for broader application and dynamic cooperation.
- Automation vs. Human Control: Explore balancing automation and human involvement to optimize multi-agent workflows and interactions.
- Efficiency Optimization: Investigate strategies to refine agent topology and conversation patterns.

ART: Automatic Reasoning and Tool-use

Paranjape et al., (2023)

- CoT prompting requires hand-crafting task-specific demonstrations and carefully scripted interleaving of model generations with tool use.
- ART is a framework that uses frozen LLMs to automatically generate intermediate reasoning steps as a program.



Comparison between related approaches

Table 1: Comparing ART with related approaches for multi-step reasoning and tool-use

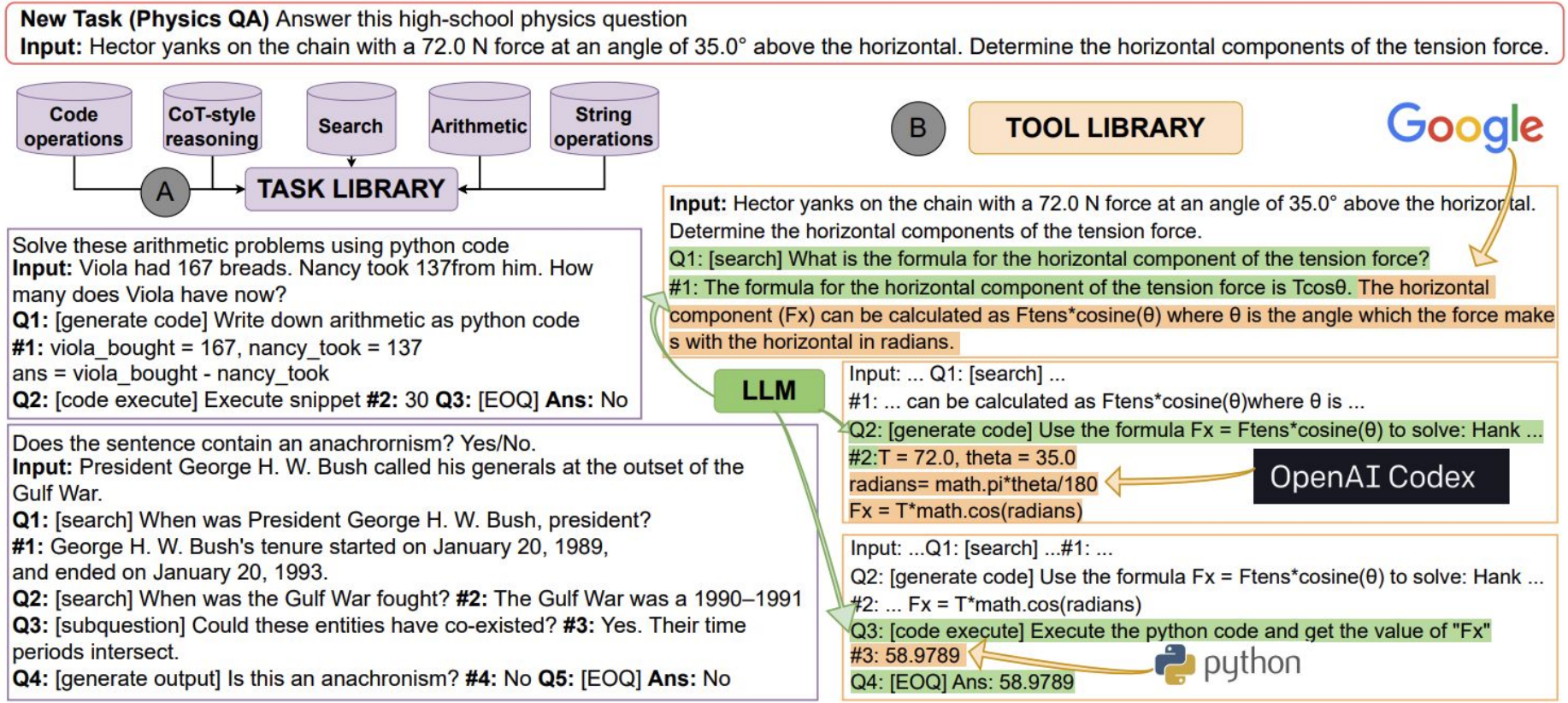
Feature	CoT	Auto CoT	Tool-former	ART
Multi-step reasoning	✓	✓		✓
Limited supervision		✓	✓	✓
Tool use			✓	✓
Extendable libraries				✓
Cross-task transfer		✓	✓	✓
Human feedback	✓			✓

Task Library

- Tasks from Big-Bench: Arithmetic, Code, Search and question decomposition, Free-form reasoning, and String Operations.
- Program grammar
  - A flexible format to accommodate a wide variety of NLP tasks.
  - Consists of a series of nodes: input node (with task name, instruction, and input), sub-step nodes (QA pairs), and a final answer node.
- Dynamic retrieval system to select relevant tasks
  - Labeled examples
  - Crafted few-shot prompts

Tool Library

- ART pauses generation when a sub-task query matches a tool name.
- Seeded Tools
  - Search Tool: SerpAPI for Google Searches. Marked as “[search]”
  - Code Generation: Codex for generating Python Code. Marked as “[generate python code]”
  - Code Execution: Virtual Python Environment


New Task (Physics QA) Answer this high-school physics question
Input: Hector yanks on the chain with a 72.0 N force at an angle of 35.0° above the horizontal. Determine the horizontal components of the tension force.
Code operations
CoT-style reasoning
Search
Arithmetic
String operations
A
TASK LIBRARY
B
TOOL LIBRARY
Google
Solve these arithmetic problems using python code
Input: Viola had 167 breads. Nancy took 137from him. How many does Viola have now?
Q1: [generate code] Write down arithmetic as python code
#1: viola_bought = 167, nancy_took = 137
ans = viola_bought - nancy_took
Q2: [code execute] Execute snippet #2: 30 Q3: [EOQ] Ans: No
Does the sentence contain an anachronism? Yes/No.
Input: President George H. W. Bush called his generals at the outset of the Gulf War.
Q1: [search] When was President George H. W. Bush, president?
#1: George H. W. Bush's tenure started on January 20, 1989, and ended on January 20, 1993.
Q2: [search] When was the Gulf War fought? #2: The Gulf War was a 1990–1991
Q3: [subquestion] Could these entities have co-existed? #3: Yes. Their time periods intersect.
Q4: [generate output] Is this an anachronism? #4: No Q5: [EOQ] Ans: No
Input: Hector yanks on the chain with a 72.0 N force at an angle of 35.0° above the horizontal. Determine the horizontal components of the tension force.
Q1: [search] What is the formula for the horizontal component of the tension force?
#1: The formula for the horizontal component of the tension force is Tcosθ. The horizontal component (Fx) can be calculated as Ftens*cosine(θ) where θ is the angle which the force makes with the horizontal in radians.
LLM
Input: ... Q1: [search] ...
#1: ... can be calculated as Ftens*cosine(θ)where θ is ...
Q2: [generate code] Use the formula Fx = Ftens*cosine(θ) to solve: Hank ...
#2:T = 72.0, theta = 35.0
radians= math.pi*theta/180
Fx = T*math.cos(radians)
OpenAI Codex
Input: ...Q1: [search] ...#1: ...
Q2: [generate code] Use the formula Fx = Ftens*cosine(θ) to solve: Hank ...
#2: ... Fx = T*math.cos(radians)
Q3: [code execute] Execute the python code and get the value of "Fx"
#3: 58.9789
Q4: [EOQ] Ans: 58.9789
python

Human Feedback

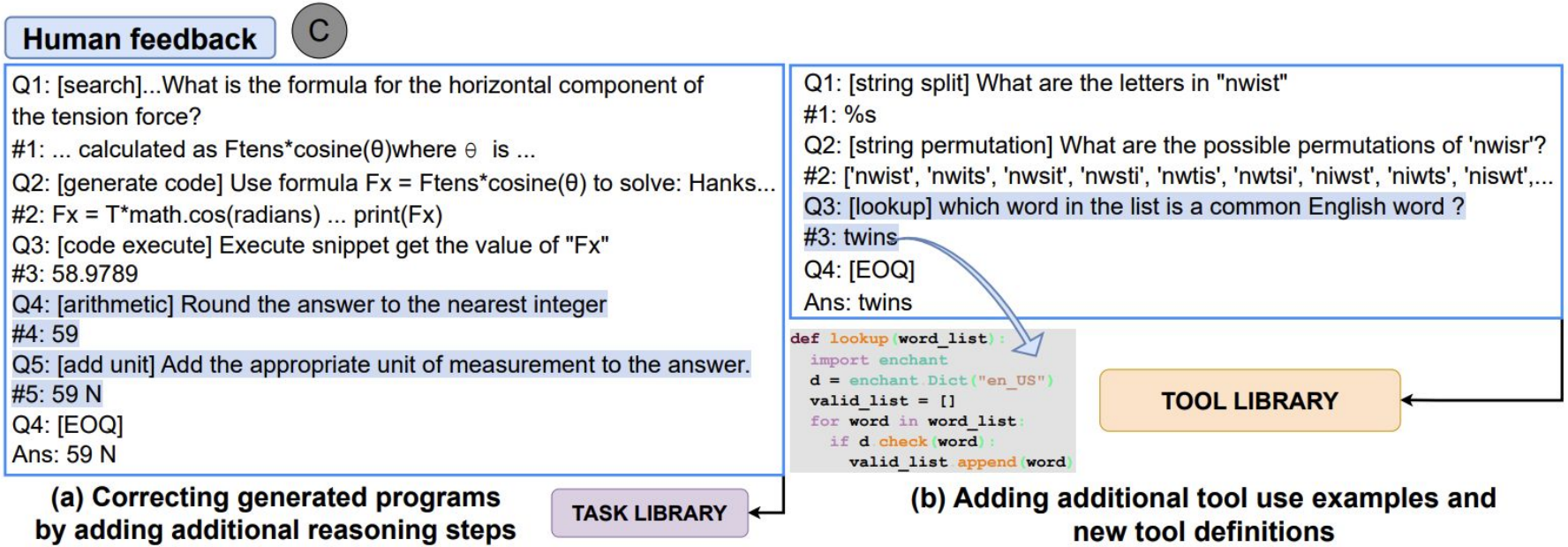


Figure 3: Human feedback to ART shown for (a) PQA where reasoning steps are added to the program and; (b) Word unscrambling where tool library is augmented with a new lookup tool.

Evaluation

- Tested on tasks from its own library, BigBench, MMLU, etc.

Task Name (Cluster)	Few Shot	AutoCot	ART w/o Tool Use	ART	GPT-3 Best
Anachronisms (Search)	71.3[5]	51.48	70.87	75.66	-
Musique (Search)	2.03[5]	12.88	10.04	19.19	15.2[3]
Hindu Knowledge (Search)	85.02 [5]	73.03	83.42	87.98	-
Known Unknown (Search)	68.90 [5]	56.09	80.43	80.43	-
Δ with ART (Search)	**+9.0**	**+17.44**	**+4.6**		+4.0
Elementary Math QA (Arithmetic)	56.40[7]	74.52	58.04	68.04	-
Aqua-rat (Arithmetic)	20.54[7]	34.41	36.29	54.20	54.1[4]
GSM8K (Arithmetic)	7.79[7]	21.99	53.4	71.00	71.6[4]
Navigate (Arithmetic)	60.7[7]	61.7	72.4	72.4	85.90[1]
Δ with ART (Arithmetic)	**+30.0**	**+18.25**	**+11.4**		-4.7
K'th letter concatenation (String)	3.2[5]	0.64	8.19	40.00	98.0[2]
Language games (String)	35.14[5]	18.58	11.19	23.08	-
Date Understanding (String)	37.53[5]	38.90	52.05	-	70.41[1]
Auto Debugging (Code)	62.94[5]	38.24	55.29	62.94	-
Code Description (Code)	97.99[7]	88.67	84.67	88.00	-
Formal Fallacies (CoT)	44.84[5]	56.4	64.76	-	58.4[1]
Hyperbation (CoT)	62.72[5]	55.4	80.80	-	72.4[1]
Δ with ART (Misc)	**+9.6**	**+16.4**	**+13.7**		-15.4
Δ with ART (Overall)	**+14.90**	**+17.17**	**+7.91**		-9.0

Evaluation

- Evaluated on test tasks without explicit supervision.

Task Name (Cluster)	Few Shot	AutoCot	ART w/o Tool Use	ART	GPT-3 Best
Δ with ART (Overall)	**+6.9**	**+24.6**	**+16.7**		-1.7

	MMLU				
College Computer Science (Search)	41.00	43.99	63.40	67.80	63.6[6]
Astronomy (Search)	62.10	41.48	76.71	79.1	62.5[6]
Business Ethics (Search)	61.60	48.8	77.17	81.16	72.7[6]
Virology (Search)	50.03	49.52	71.60	71.49	50.72[6]
Geography (Search)	77.67	57.07	70.30	71.71	81.8[6]
Mathematics (Arithmetic)	36.67	33.77	39.50	45.66	34.5[6]
Δ with ART (MMLU)	**+14.6**	**+23.7**	**+3.0**		**+8.5**

Future directions for ART

- Implement Self-Consistency
  - Utilize generating multiple outputs for the same task and selecting the most common solution to enhance accuracy and reasoning depth.
- Enhance through human feedback
  - Integrate manual corrections and additions to programs based on task-specific feedback to rectify errors and enrich the tool and task libraries.
- Expand libraries with quality demo
  - Systematically incorporate corrected programs and new tools from human feedback into ART's libraries to boost its adaptability and effectiveness across a broader range of tasks.

Q & A

- Prevention of risky agent conversations leading to negative consequences
  - Built-in safety protocols to avoid engaging in harmful or unethical behaviors.
  - Human-in-the-loop to ensure human oversight to intervene.
- The program may have errors, such as package non-existence, data type error, etc. How does AutoGen correct errors and rewrite?
  - Handle by assistant agent.
- In dynamic group chat, how LLM decide the next speaker among agents with different expertise?
  - Relevance Assessment: Evaluate each agent's domain expertise
  - Contribution Potential: Provide most valuable information
  - Human Feedback