



SocioDojo: Building Lifelong Analytical Agents with Real-world Text and Time Series

Junyan Cheng, Peter Chin

Thayer School of Engineering, Dartmouth College

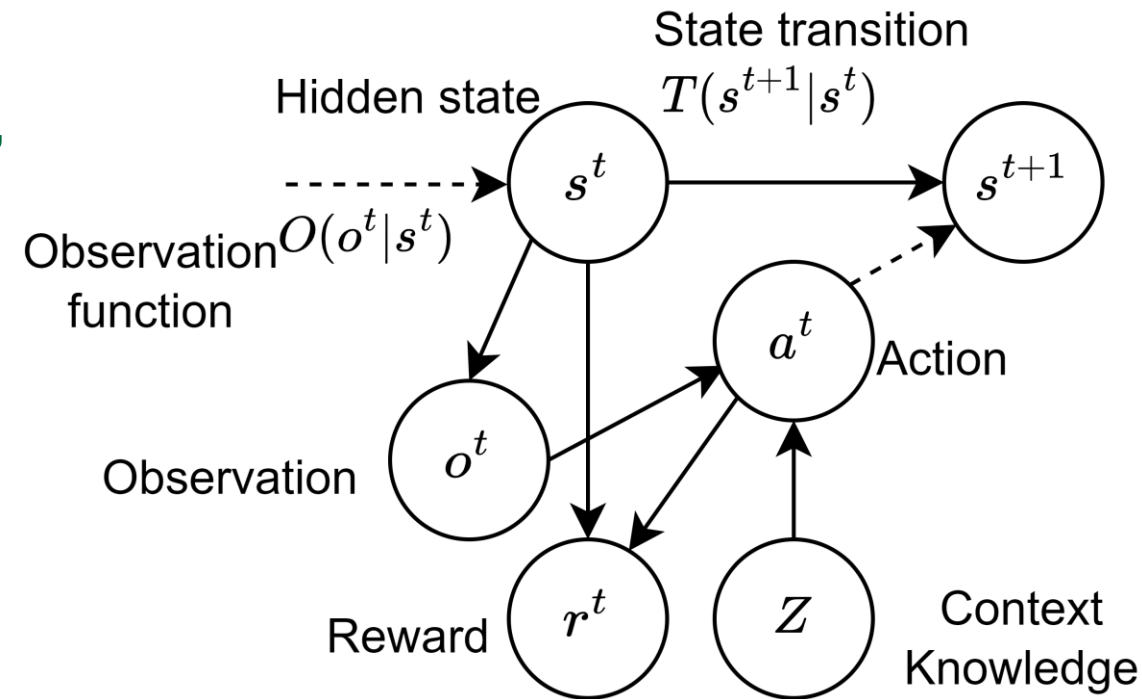
ICLR 2024

- **Objective:** Promoting the agents that can comprehend, analyze, and make decisions in real human society.
- **Existing methods:**
 - Games: Human society is too complex to emulate.
 - Exams: “expert” extensive; hard to include up-to-date knowledge.
- **Design targets:**
 - Assess the societal analysis and decision-making capabilities of agents reliably.
 - Keep up with the evolution of society over time in a low-cost way.
- **Key insights:**
 - Real-world time series are probing the hidden status s^t of the society.
 - “Society model” $P(s^{t'} | \omega^t)$ given observation ω^t provides basis for real-society works.

Hyperportfolio Task



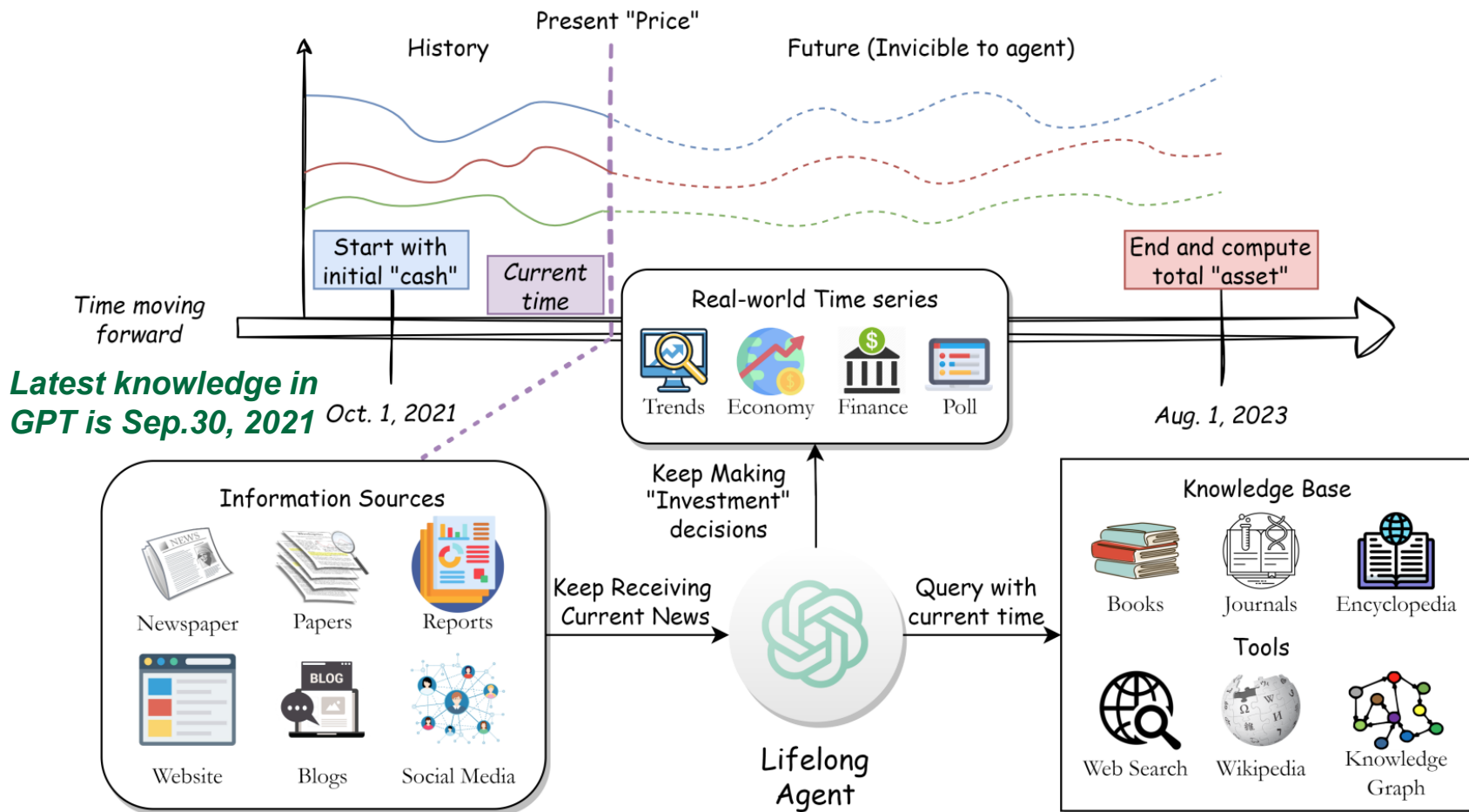
- Precisely modeling the state is non-trivial, we apply a surrogate decision-making task “Hyperportfolio” which takes time-series not restricted to finance as assets.
- Agent manages a hyperportfolio over time denoted as POMDP $(O, T, S, A, \Omega, R, Z)$:
 - **Observation** $o \in \Omega$ in society are news, articles, internet posts, reports, papers, etc.
 - **Observation function** O are real-world media, social network, research institutes, etc.
 - **State** s and **transition function** T are hidden, probed by time series, e.g., national statistics.
 - **Context** Z comes from books, experience, etc.
 - **Actions** $a \in A$ are buy, sell, or hold assets.
 - **Rewards** $r \in R$ is the expected return.



SocioDojo Environment



The agent starts at an initial date with initial cash, as time moves forward, it will continuously receive the latest news and take actions to adjust its portfolio until the end date. The total return will be settled on the end date as the score of the agent.

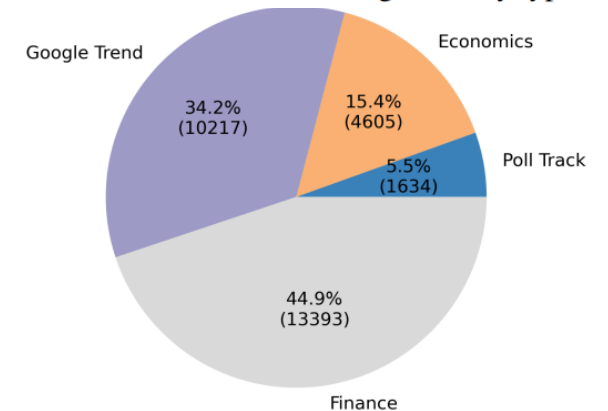


Type	Num	Avg. Tok.	Per day
Newspapers	175.2K	674	262
Social media	259.8K	50	388
Reports	19.5K	4926	29
Papers	6.2K	26939	9
Other	32.7K	1596	49

Table 1: Statistics of information sources by type.

Type	Keys	Tokens	Avg. Tok.
Encyclopedia	57.4K	14.8M	258
Books	167.5K	41.0M	245
Papers	1.7M	392.4M	228

Table 2: Statistics of knowledge base by type.

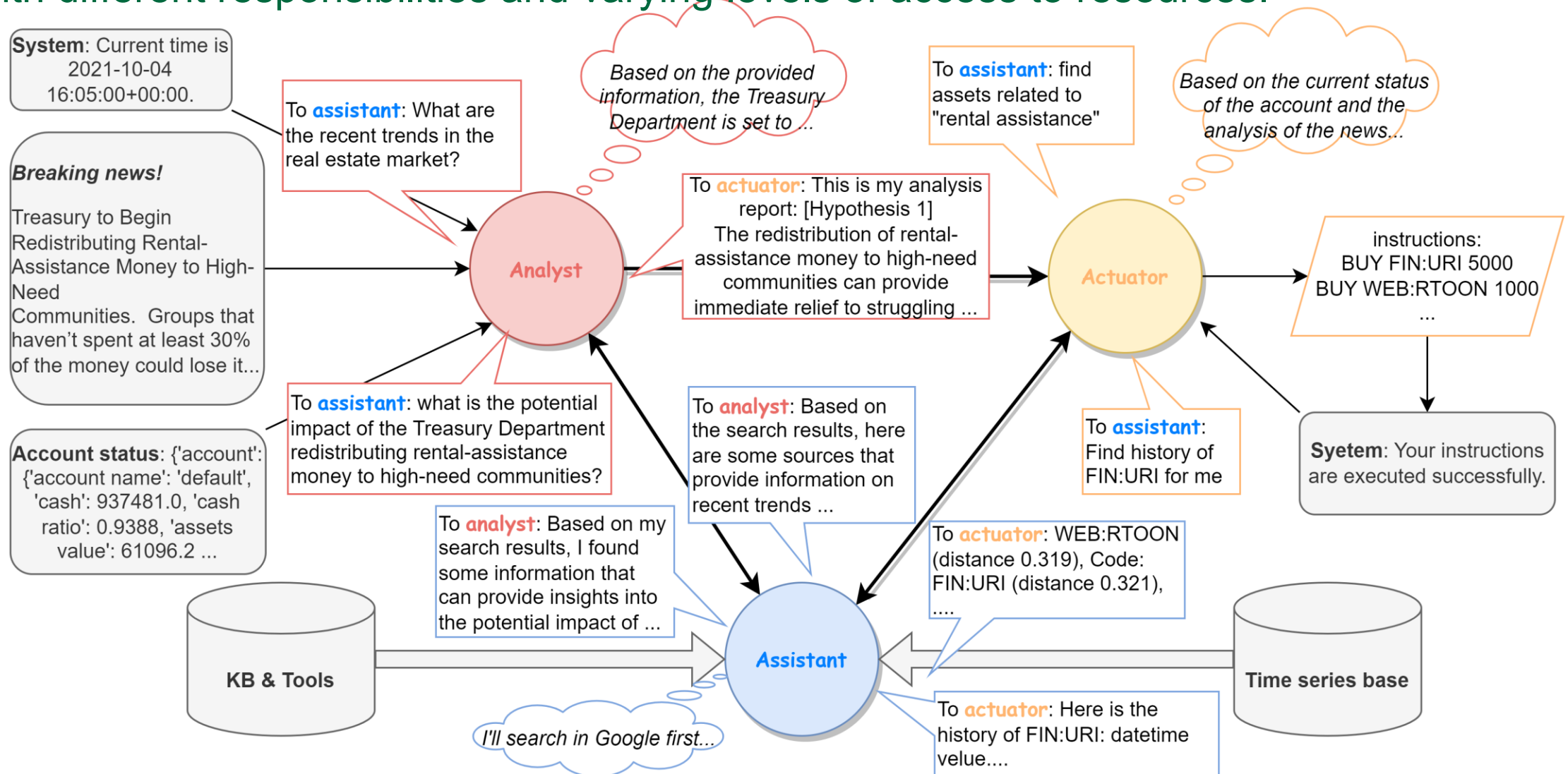


- **Objective:** Improvements in metric should stem from analytical prowess, not luck
- **Mechanisms Inspired by Financial Markets:**
 - **No Day-Trade Rule:** the asset can be sold after 5 days of buying in order to avoid the “front run” as our financial data is only daily *close prices*.
 - **Commission Fee:** around 1% of trading value for each trade to avoid frequent operations.
 - **Overnight Fee:** charge a portion of the current value at the end of every day for some categories of time series to avoid speculation and some time series are ever-growing.
 - **Return Bound:** the maximal return or loss from an asset is bounded by X times the initial investment amount.
- Determining the mechanism parameters by minimizing expected return from undesirable behaviors (“luck”) $E(R_u)$ modeled by a random walk agent.

Analyst-Assistant-Actuator Architecture



AAA Agent Architecture: An architecture that encompasses three distinct roles, each with different responsibilities and varying levels of access to resources.



Hypothesis & Proof Prompting



- Hypothesis & Proof prompting, input to the analyst as part of the role context:

In order to make a convincing analysis, you must clearly state your hypothesis and provide evidence to support your hypothesis:

1. If your hypothesis cannot be supported by the facts, knowledge, and information you have, you should seek the help of another assistant agent on your team by calling the `ask` function.
2. You should never make up facts, knowledge, or information, if you do not know, you should call the `ask` function to seek help from the assistant agent.
3. The assistant agent will find evidence to support your hypothesis if you call the `ask` function and send the evidence to you. It can also search for general information, facts, and knowledge, as well as information on an ICode.
4. The hypothesis can also be rejected by the evidence; in that case, you should give up the hypothesis and construct a new one.
5. Your hypothesis may lack support for both acceptance and rejection; in that case, you can choose to insist on it as an intuition, but you should explain why you insist on it, or you can choose to give up it.
6. When you are done, you should call the `done` function to finish the analysis.
7. You must call one function `done` or `ask` in each step of the analysis.

Here are some tips for you to make a hypothesis:

1. A hypothesis is like a guess, assumption, or intuition, it is a statement that you think is true, but you do not have enough evidence to support it yet.
2. For example, "interest increase in technology will cause the technology stock to increase", "sale of lipstick indicates a potential drop of the economy", "the president's approval rate will increase if the economy is good", etc.
3. The hypothesis can be microscopic or macroscopic, it can be about a specific asset or a general trend.

Now, progressively analyze the given news, article, report, etc. in a multi-round dialog between you and the `ask` function. When you think that you get enough information to give the final analysis report, call the `done` function to end the dialog with the `ask` function, and aggregate the progress as a final analysis report when the system asks you to give your final analysis report. You must call one of the `done` or `ask` functions in your response.

- **Plan & Solve (Wang et al., ACL, 2023):** “Let's first understand the news, article, report, etc. and devise a plan to analyze the news, article, report, etc. Then, let's carry out the plan to analyze the news, article, report, etc. step by step. You can call `ask` function to ask questions to another assistant agent for help or search for the evidence, facts, knowledge, information, or data demanded by you.”

Comparison with Plan & Solve



- Plan & Solve analyst asks only one simple fact-check question:

[Analyst] Ask assistant: **What is the Epic antitrust case against Apple about?**

[Assistant] The Epic antitrust case against Apple, officially known as Epic Games v. Apple, is a lawsuit brought by Epic Games against Apple in August 2020. The case was filed in the United States District Court for the Northern District of California. According to the search results from the wikisearch, Epic Games filed the lawsuit against Apple for antitrust violations and anti-competitive behavior, along with a similar lawsuit against Google. The lawsuit alleges that Apple's app store practices violate antitrust laws. More information about the case can be found in the search results.

[Analyst] Thank you for the information. Based on the news article, it seems that Apple is planning to appeal the verdict in the Epic antitrust case. The article mentions that Apple wants to stay on the judge's order to open in-app communications that could allow cheaper payment options. The case focuses on how big technology companies, like Apple, control access to their platforms and the restrictions they impose on app developers... (omitted here)

I'm ready to give the final analysis now.

- Hypothesis & Proof analyst asks 3 iterative deeper in-depth questions:

1. **What is the potential impact of Apple filing notice of appeal in the Epic antitrust case?**
2. **Has there been any significant market reaction to the news of Apple filing an appeal in the Epic Games antitrust case?**
3. **Is there any evidence of market reactions to legal proceedings or appeals in similar antitrust cases involving large technology companies?**

Experiment setting



- **Two task settings:**

- **Standard:** Agents have the freedom to select any assets to construct their portfolio.
- **Tracking:** Agents can only opt for assets from a curated list of 100 representative time series. Examples include selections from Dow Jones 30 stocks.

- **Prompting and LLM agent baselines:**

Method	Type	Strategy	Application
Plan & Solve (Wang et al., ACL, 2023)	Zero-shot prompting	Instructs the agent to devise a problem-solving plan and subsequently implement it.	Paired with the analyst in our experiment.
Self Ask (Press et al., arXiv 2210.03)	Few-shot prompting	Encourages the agent to pose relevant queries about ambiguous portions when responding.	Paired with the assistant in our experiment.
ReAct (Yao et al., ICLR, 2023)	Few-shot prompting	Prompts the agent to modify its response or action based on the observed feedback.	Paired with the assistant in our experiment.
AutoGPT (Richards, 2023)	Autonomous LLM agent framework	We use it to produce an analysis report for an input and pass to actuator.	Play the role of the analyst in our experiment.

- We use **GPT-3.5-Turbo** by default with **temperature=0.2** for more deterministic results for reproduction while preserving some creativity for better performance.

Experiment results



	H&P (Ours)		P&S ¹	Self-Ask ²	ReAct ³	AutoGPT ⁴	Vanilla		
	full	w/o AAA					w/ Ask	w/o Ask	
S.	Fin	15.3	13.0	12.4	11.8	10.9	9.8	10.4	6.1
	Econ	5.5	4.5	1.0	3.6	2.2	-6.3	1.8	0.2
	Web	15.0	13.8	11.1	9.4	9.5	8.3	10.2	-4.5
	Poll	3.6	2.2	4.1	-2.9	1.7	3.1	-1.2	2.5
	All	14.3	11.9	10.8	9.9	9.4	8.9	9.8	3.1
<i>S</i>	3.21	2.91	2.90	2.60	2.15	2.77	2.16	1.20	
T.	Fin	15.6	13.9	12.0	11.6	10.3	8.6	10.7	8.0
	Econ	2.8	-4.2	-3.2	-1.1	-1.7	1.0	N/A	0.2
	Web	13.7	14.9	8.7	4.3	4.9	13.3	3.8	7.3
	Poll	3.1	2.7	N/A	1.4	-1.1	2.3	0.1	-6.7
	All	15.0	13.1	11.5	10.6	8.8	10.3	10.0	6.6
<i>S</i>	3.07	2.42	2.68	2.15	2.43	2.45	2.35	1.67	

**Numbers in the table denote percentage returns at the end time.*

- Both P&S and our method enhance analysts, leading to improvements.
 - The quality of analysis is paramount and serves as the linchpin for the entire system.
- Our approach prompts more questions with better incisive thereby elevating the quality of analysis.
 - e.g., our method poses: “What is the potential impact of X on Y?” compared to others simply ask: “What is X?”.
- Absent the assistant (“Ask”), models underperform due to their lack of access to recent knowledge.
- Without the AAA architecture, the performance drop may be due to conflicts of roles.

	H&P	P&S	Other
Mean	1.91	0.72	0.52
Std.	0.98	0.88	0.85

Table 4: Number of ask calls made by agents in our experiments.

• Ablation Study: Information Sources

- Challenges exist in fully utilizing certain sources:
 - Lack of multimodal processing abilities (e.g., visual languages).
 - Token restrictions cause content like papers to be fragmented.
- The quality and reliability of sources matter, with platforms like social networks potentially introducing noise.

Type	Fin	Econ	Web	Poll	All
News.	15.3	5.5	15.0	3.6	14.3
+Social.	14.3	5.3	15.3	2.3	14.1
+Report	15.9	6.3	14.2	3.9	14.7
+Papers	15.3	6.7	15.0	3.4	14.6
+Other	15.0	5.3	14.9	4.9	14.4
Mixed	15.5	5.8	15.4	4.0	14.7

Table 5: Ablation study of types of sources.

• Ablation Study: Foundation Models

- GPT-4 shows overall enhancement which is aligned with its improvement in social science tests ([OpenAI, arXiv 2303.08](#)).
- Better analysis from H&P alleviates the bottleneck faced by the Assistant and Actuator comes from the Analyst.

Model	Fin	Econ	Web	Poll	All
GPT-3.5	15.3	5.5	15.0	3.6	14.3
GPT-4	17.2	7.3	16.7	4.2	16.8
Assis.	15.9	5.8	16.4	4.3	15.2
Actua.	15.8	6.0	15.3	3.1	14.9
Analy.	16.9	6.8	16.2	4.4	15.9

Table 6: Ablation study on foundation models by replacing “GPT-4” for all roles or for each of “Assistant”, “Actuator”, “Analyst”.



Thank you!