Generative AI for reliability engineering

Analyze data

Certified Reliability Engineer exam of American Society of Quality (ASQ)

- Short question with poor context -> search agent?
- Many problems involving calculations -> code agent?

-> Multi-agent LLM system

Agentic Reasoning: Reasoning LLMs with Tools for the Deep Research

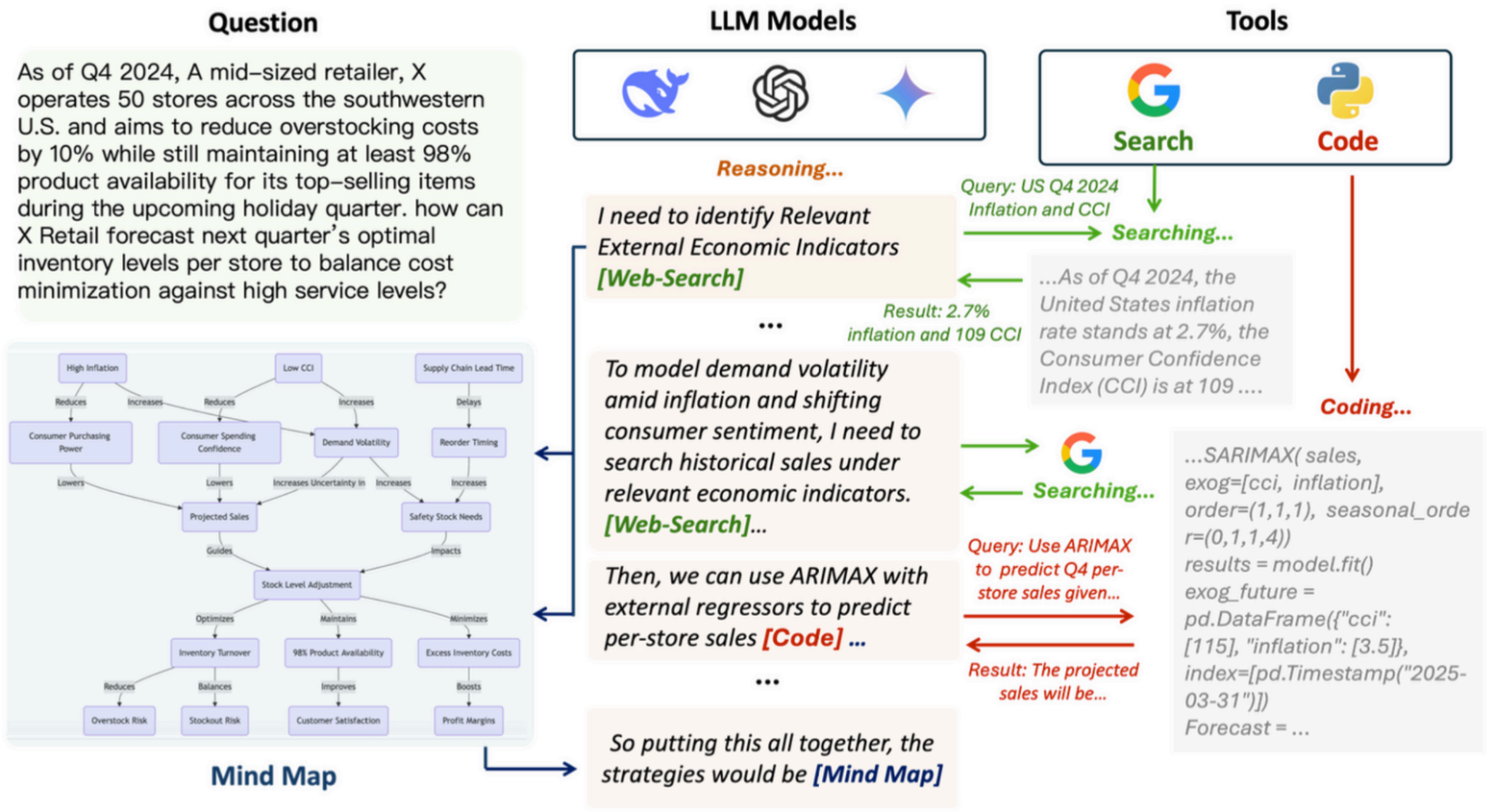
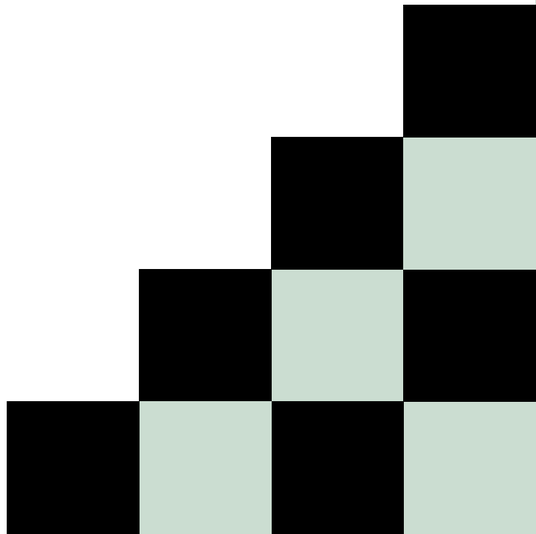


Figure 1: The overall workflow of Agentic Reasoning.

Paper implementation based on actual code

Why?

- Too much code, difficult to modify
- Too many dependencies for the same usage
- Bad result for this data



Paper implementation based on actual code

- GPT-4o as basic model
- Tavily as search agent
- liteLLM for calling all of models
- nano-graph for mind map

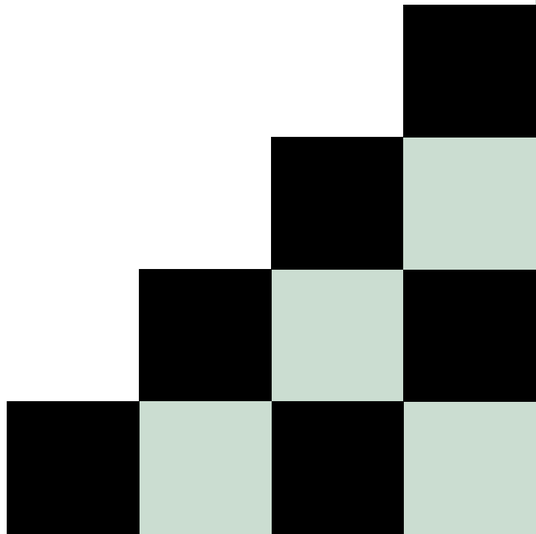
Result

	Accuracy on train data	Accuracy on test data (public)	Accuracy on test data (private)
GPT-4o	0.64	0.50	0.42
GPT-4o-mini	-	0.75	0.55
Agentic Reasoning	0.64	0.50	0.58
Agentic Reasoning V2	**0.72**	**0.75**	**0.75**

Analyze the number of times the agent is called

- Called code agent 0.8 time per question
- Called search agent 0 time
- Called mind-map agent 0 time

-> Code agent capability is essential for the final result
-> Does forcing the search/mind map yield better results?



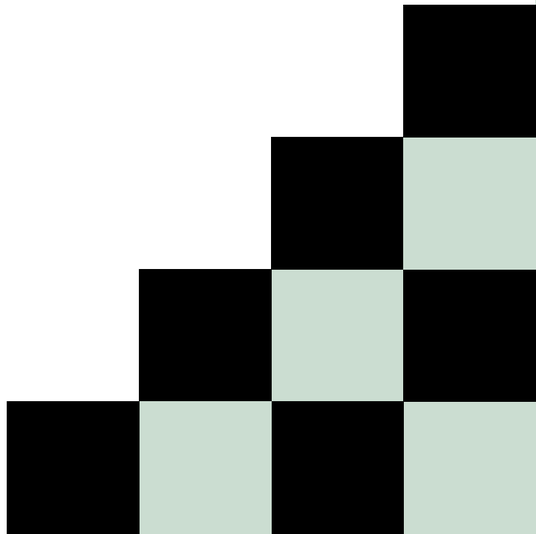
Result

	Accuracy on train data	Accuracy on test data (public)	Accuracy on test data (private)
GPT-4o	0.64	0.50	0.42
GPT-4o-mini	-	0.75	0.55
Agentic Reasoning	0.64	0.50	0.58
Agentic Reasoning V2	**0.72**	0.75	0.75
Agentic Reasoning V2 with searching context	0.68	0.66	0.83
Agentic Reasoning V2 with searching context and mind map	**0.72**	**0.83**	**0.91**

Analyze the number of times the agent is called

- Called code agent 0.64 time per question
- Called search agent 1 time per question
- Called mind-map agent 0 time

-> Code agent capability is essential for the final result



Execution Time

Only code agent takes
~ **12 minutes** for 25 query in train dataset

Code agent +search agent
~ **15 minutes** for 25 query in train dataset

Code agent + search agent + mind map agent
~ **30 minutes** for 25 query of train
~ **45 minutes** for 24 query of test

Future work

1. Try a different approach to building the agent system:
    a. Add a planning agent.
    b. Replace the mind map with a different data structure.

2. Experiment with different tools/models for each agent:
    a. Use Claude for the code agent.
    b. Use Hugging Face's Smol Open Deep Search for the search agent.

3. Prompt Engineering

Thanks!

Yulin SHI