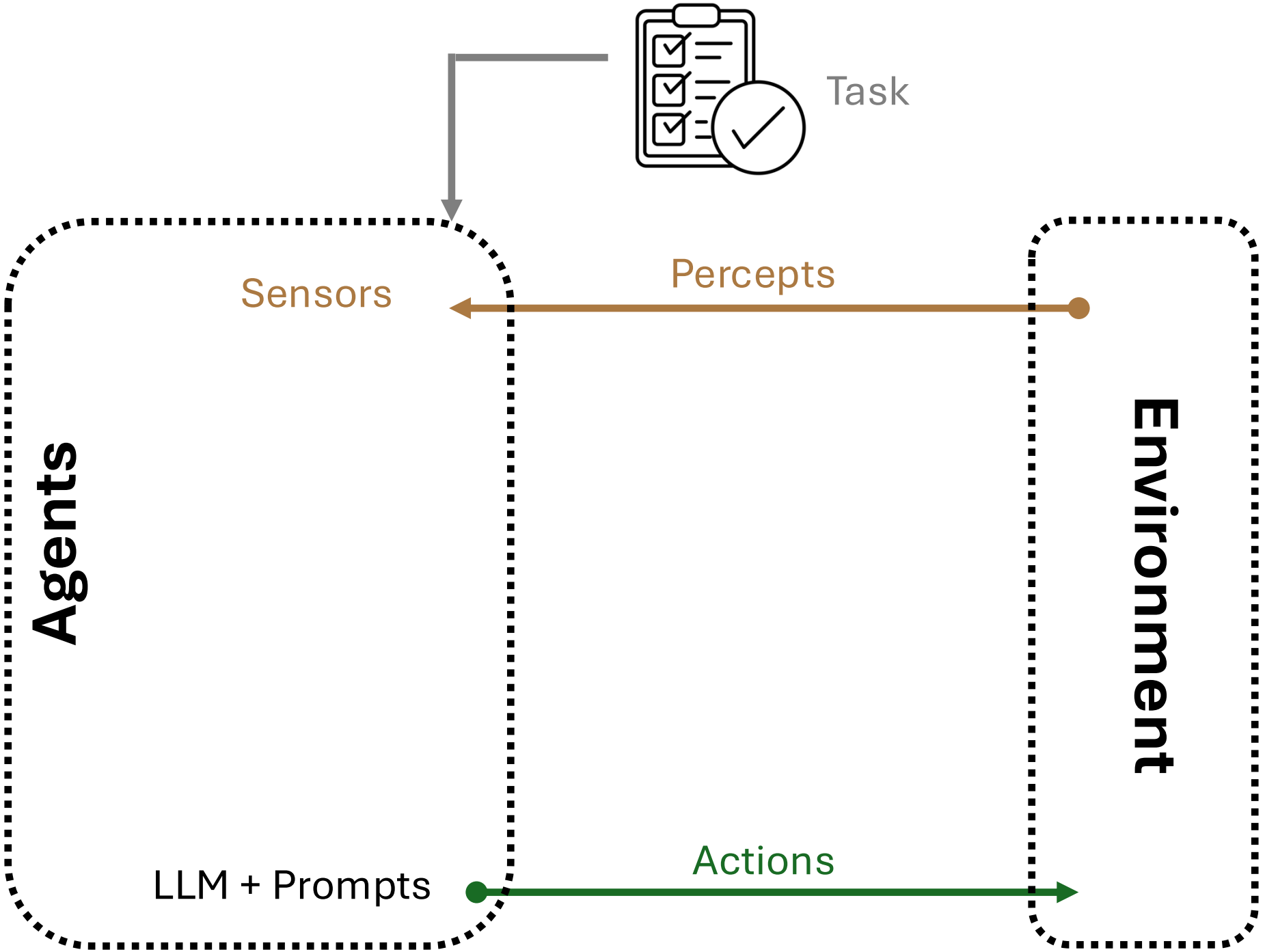
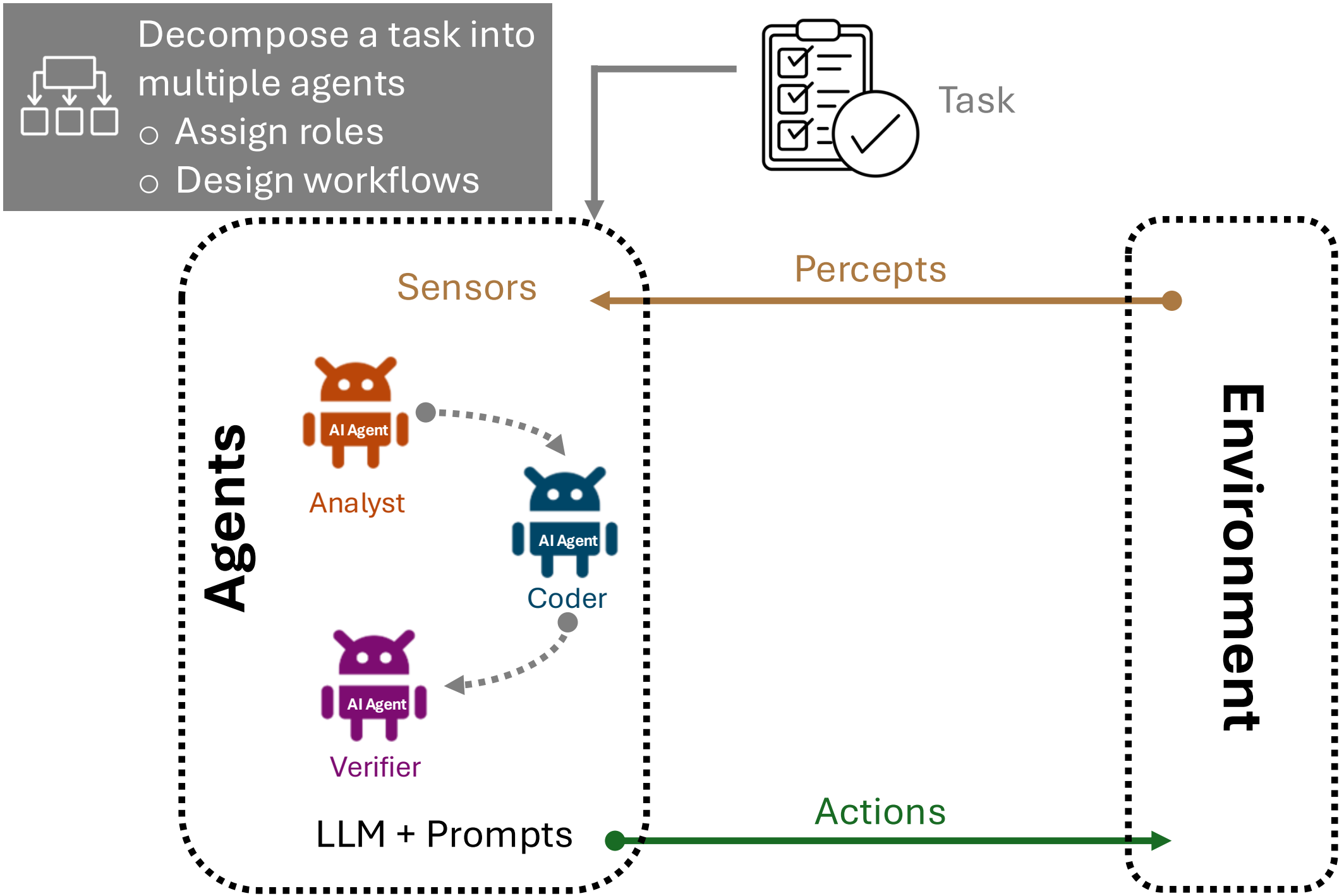
Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow

Furong Huang | https://furong-huang.com/

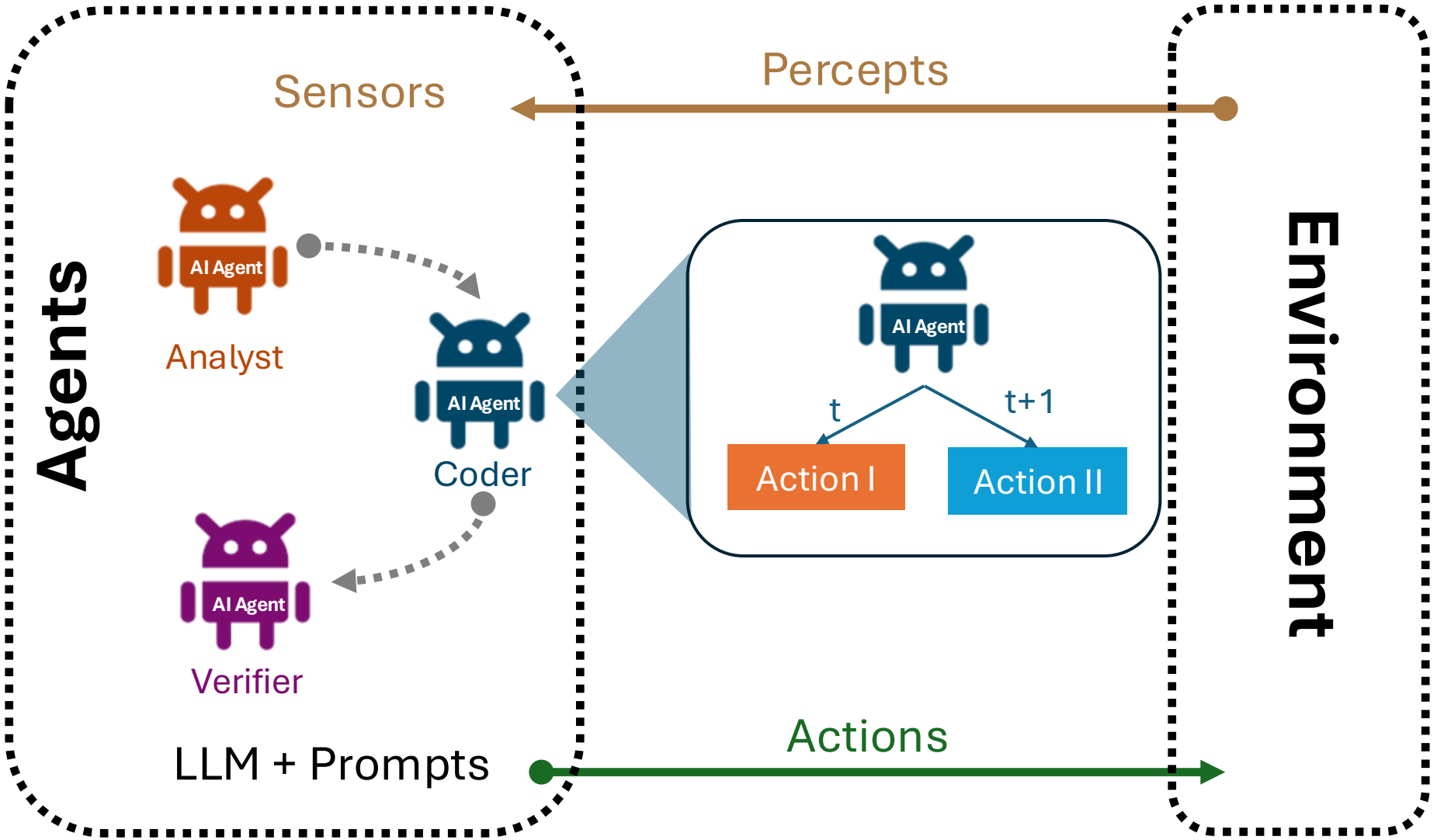
Math & AI Summer in Seattle


Task
Agents
Sensors
Percepts
Environment
LLM + Prompts
Actions


Decompose a task into multiple agents
Assign roles
Design workflows
Task
Agents
Sensors
Percepts
AI Agent
Analyst
AI Agent
Coder
AI Agent
Verifier
LLM + Prompts
Actions
Environment

Decompose a task into multiple agents
- Assign roles
- Design workflows

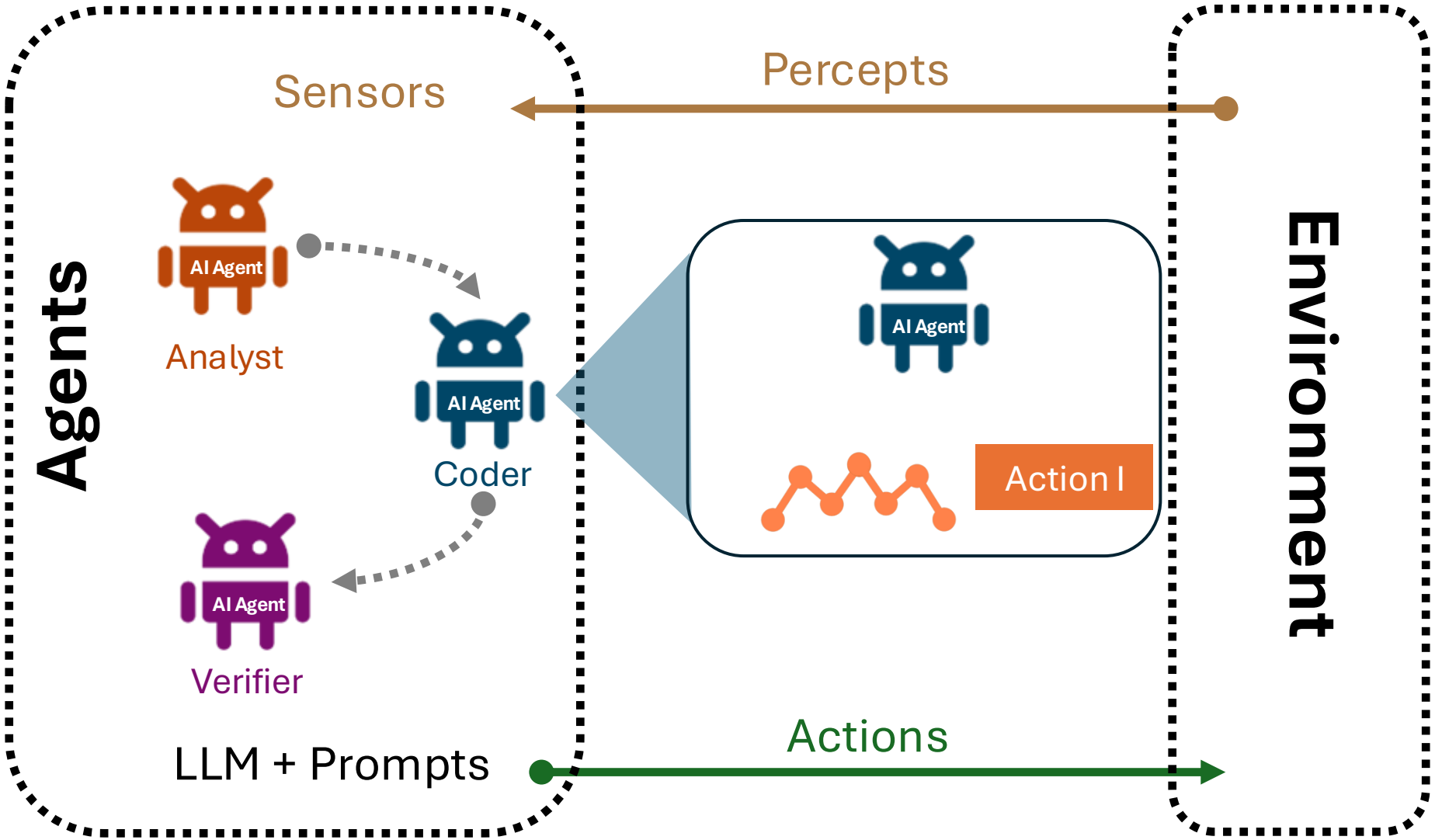
Each agent decides the action at every time step.

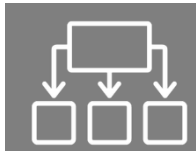


Decompose a task into multiple agents
- Assign roles
- Design workflows

Each agent decides the action at every time step.

For each action decision, the agent generates a reasoning trace.

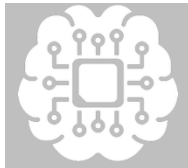




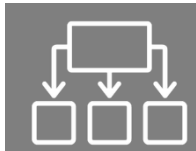
Decompose a task into multiple agents

- Assign roles
- Design workflows

Each agent decides the action at every time step.



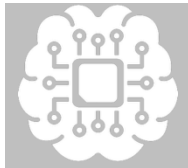
For each action decision, the agent generates a reasoning trace.



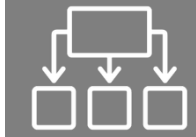
Decompose a task into multiple agents

- Assign roles
- Design workflows

Each agent decides the action at every time step.



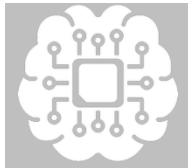
Thinking Trace/Token Control



Decompose a task into multiple agents

- Assign roles
- Design workflows

Action Control

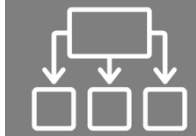


Thinking Trace/Token Control

Toward Self-Improving Agentic Systems

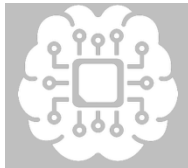
Across Thinking, Actions, and Workflow

Precompute and Reuse



Workflow Control

Action Control



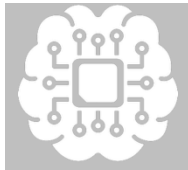
Thinking Trace/Token Control

Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow

Workflow Control

Action Control



Thinking Trace/Token Control

Alignment via Guided Controlled Decoding

$$\log \pi_{\text{decode}} = \text{Base LLM} + \text{Reward}$$

Reward $Q^*(\mathbf{s}_t, y_t)$

$$= \max_{\pi} \mathbb{E}_{\tau \sim \rho^*(\cdot|\mathbf{s}_t, y_t)} [r([\mathbf{x}, \mathbf{y}_{<t}, y_t], \tau)]$$

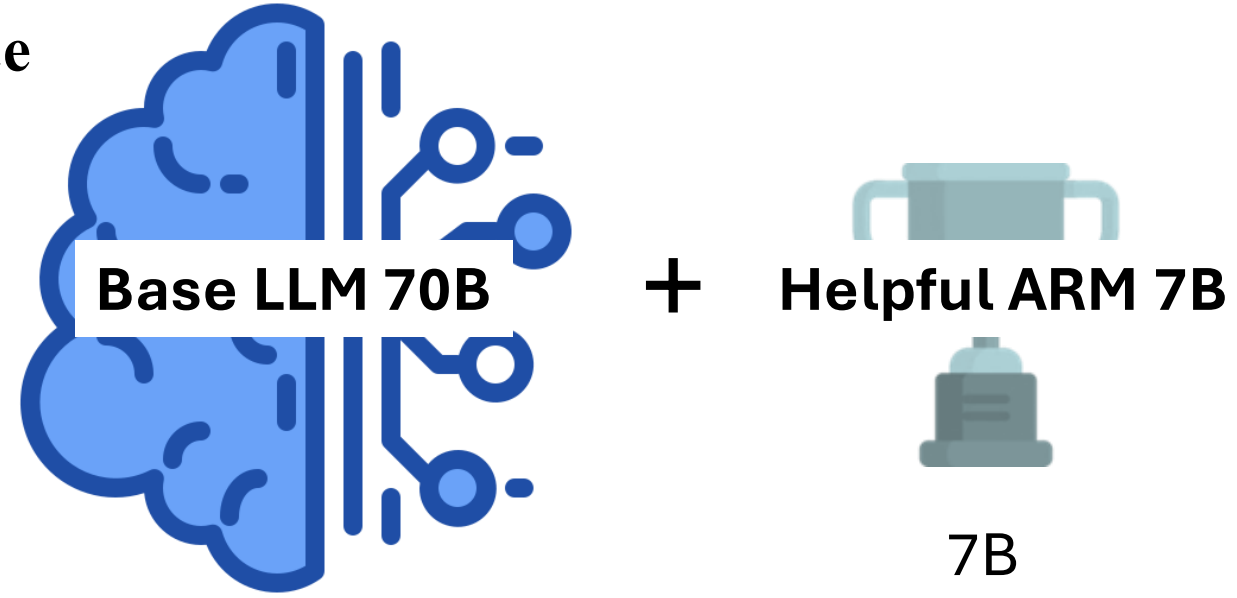
Trajectory reward under optimal policy

$$\log \pi_{\text{decode}}(y|x) = -\log Z(x) + \log \pi_{\text{base}}(y|x) + \frac{1}{\beta} r(x, y)$$

Precomputed and reusable

Exp 1: Aligning with general human preference

Exp 2: Weak-to-strong guidance



Exp 3: Multi-objective alignment (diverse preferences)

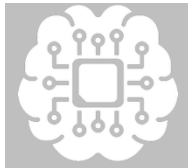
XSKZAHG, GenARM: Reward Guided Generation with Autoregressive Reward Model for Test-time Alignment, ICLR 2025.

Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow

Workflow Control

Action Control



Thinking Trace/Token Control

Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow

Imitation Learning: The “Stuck” Loop

Agent’s Faulty Plan (Imitation Learning):

1. Pick up cloth.
2. Go to sink.
3. Clean cloth.
4. Go to cabinet.
5. **Put cloth in cabinet.**

Step 9: Put cloth in cabinet.

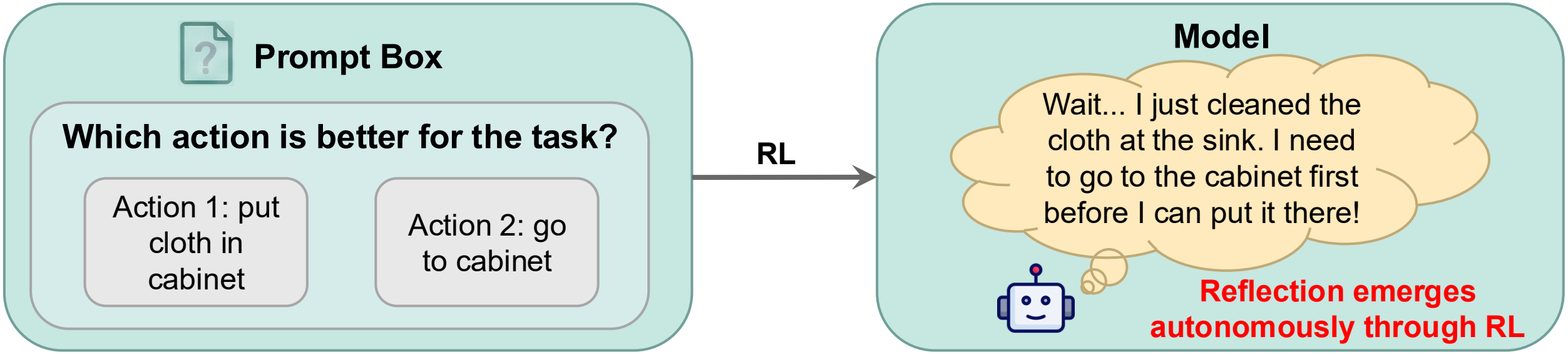
❌ **FAILS** ⚠

Step 10: Put cloth in cabinet.

Step 11: Put cloth in cabinet.

Model repeats failed action for 30+ steps until termination.

ACT: Learning Reusable ”World-Model” Knowledge



Method	ALFWorld		WebShop	ScienceWorld
	ID	OOD		
Prompt w/o CoT thinking	35.71	27.61	2.80	28.01
Prompt w/ CoT thinking	56.43	50.00	3.00	25.21
ACT	72.86	72.39	7.40	26.71
Imitation Learning	85.71	82.84	28.00	42.80
Early Experience (Self-Reflection)	87.86	85.82	31.00	45.60
IL w/ ACT	91.43	87.31	31.60	48.69
RL	90.71	84.33	29.40	43.04
RL w/ ACT	**92.86**	**88.06**	**33.80**	**50.34**

Main results on Qwen3-8B (%). ALFWorld and WebShop report success rates; ScienceWorld reports next-action prediction accuracy.

World Model Knowledge Accumulation:
RL-driven genuine self-reflection > imitating pre-generated reflection text

Generalization to General Reasoning

Model trained on ALFWorld agentic data → MATH-500, GPQA-Diamond

Method	MATH-500	GPQA-Diamond
Prompt w/o CoT thinking	78.6±0.33	42.93±1.09
Prompt w/ CoT thinking	86.93±0.74	51.52±1.89
Imitation Learning	87±0.33	44.61±0.95
Early Experience (Self-Reflection)	86.86±0.25	51.85±0.63
RL	87.07±0.77	52.36±1.32
ACT	**87.73±0.19**	**53.37±0.63**

Enhanced critical thinking
can enhance
general reasoning capabilities

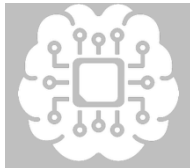
Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow

Workflow Control

Action Control

ACT trains the agent to judge *which **action** is actually better*.

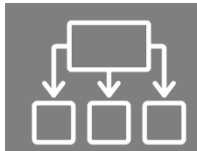


Thinking Trace/Token Control

Controlled decoding steers *how the agent **thinks***;

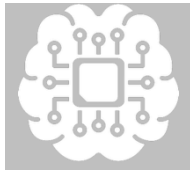
Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow



Workflow Control

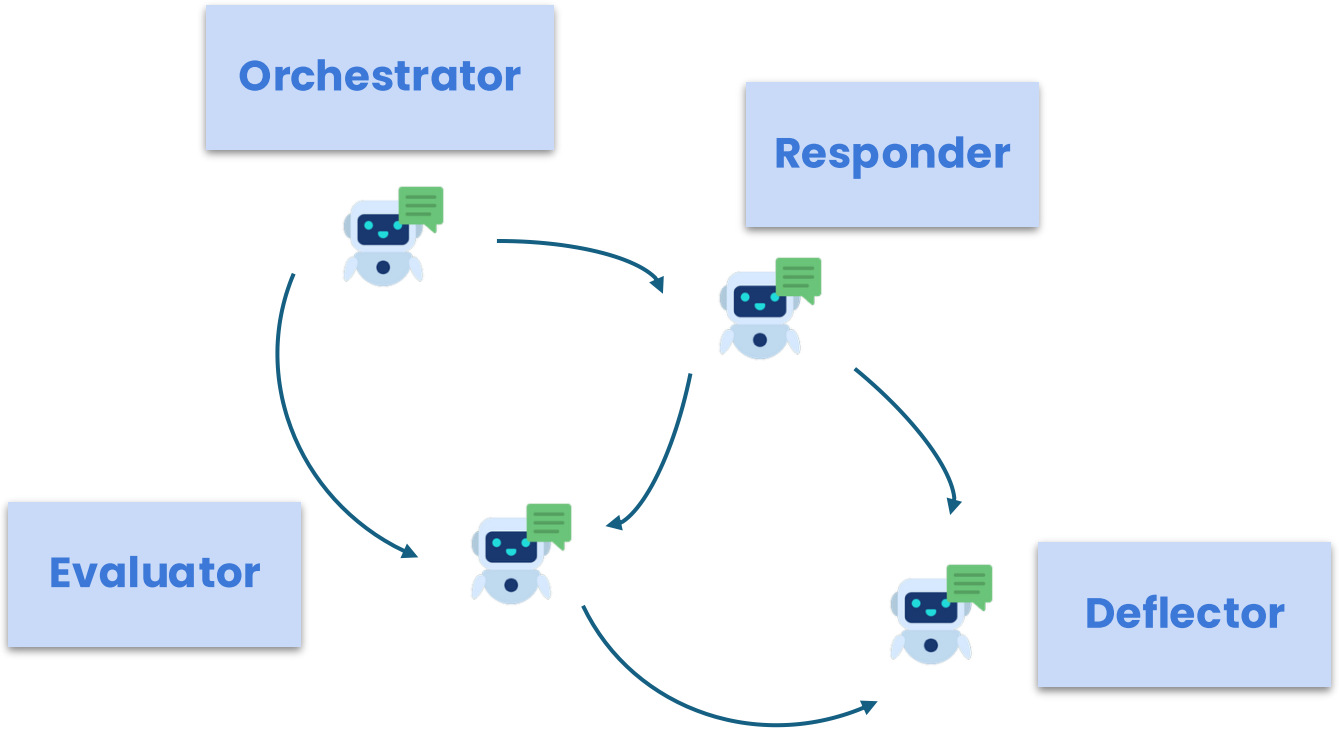
Action Control



Thinking Trace/Token Control

Agentic Workflow and Autonomous Design

Agentic Workflow: a graph of LLM calls to answer the query

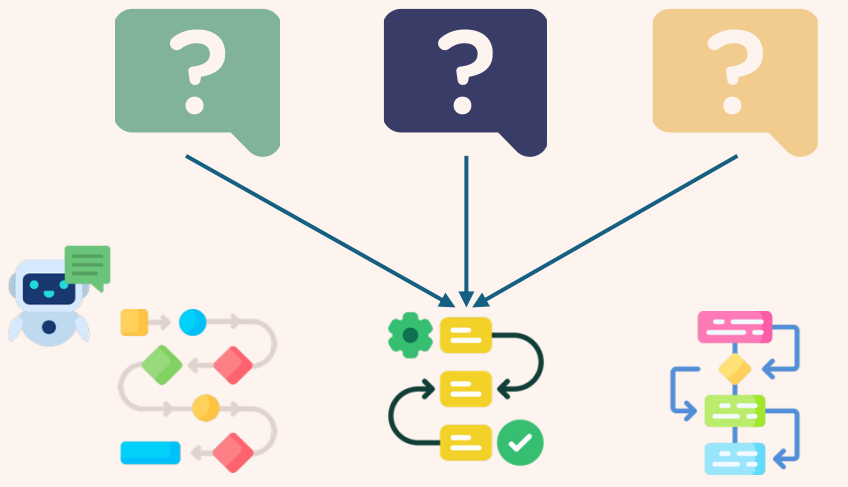


Computation Graph

Two Optimization Paradigms: One for All v.s. One for Each

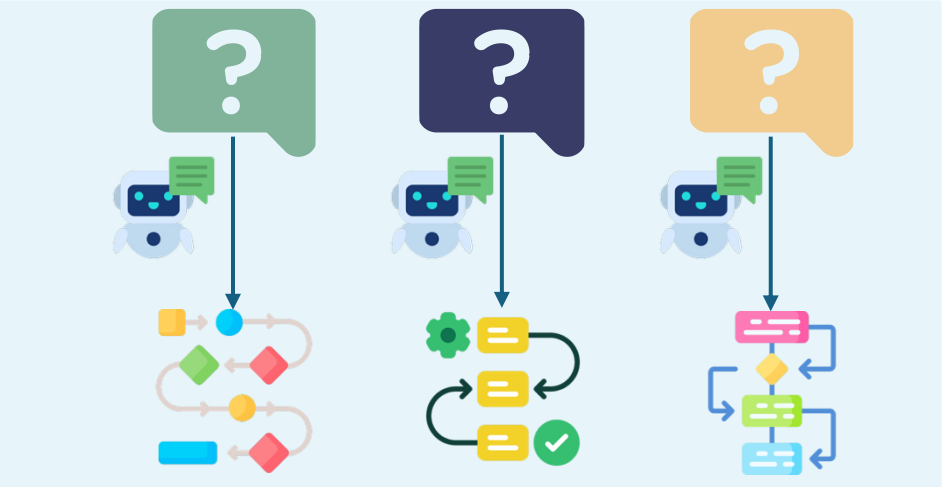
One for All

One Workflow for All Queries



One for Each

Dynamic Workflow for Each Query

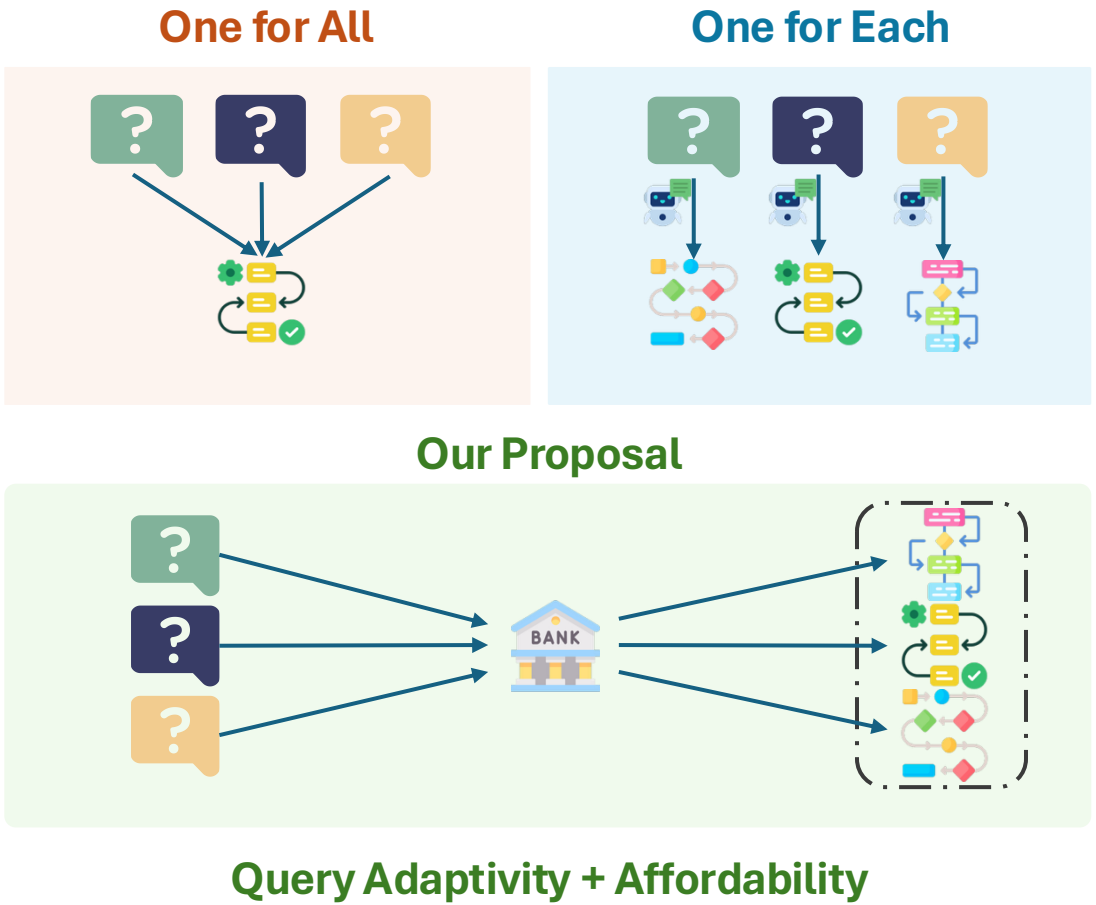


Lack Per-Query Adaptivity

Expensive to Learn

Can We Precompute Workflows in Training and Reuse in Deployment?

Can We Precompute Workflows in Training and Reuse in Deployment?



Challenges Solved:

(1) How to construct a bank?

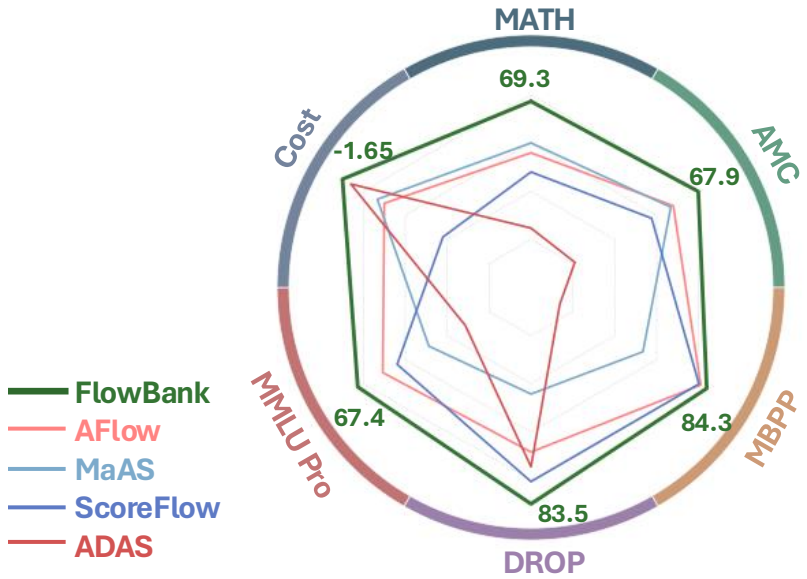
(2) How to choose during deployment?

YDYCRH, FlowBank: Query-Adaptive Agentic Workflows Optimization through Precompute-and-Reuse, arXiv:2606.11290.

Experiments – Main Results

Compared to task-level/query-level methods over 5 benchmarks

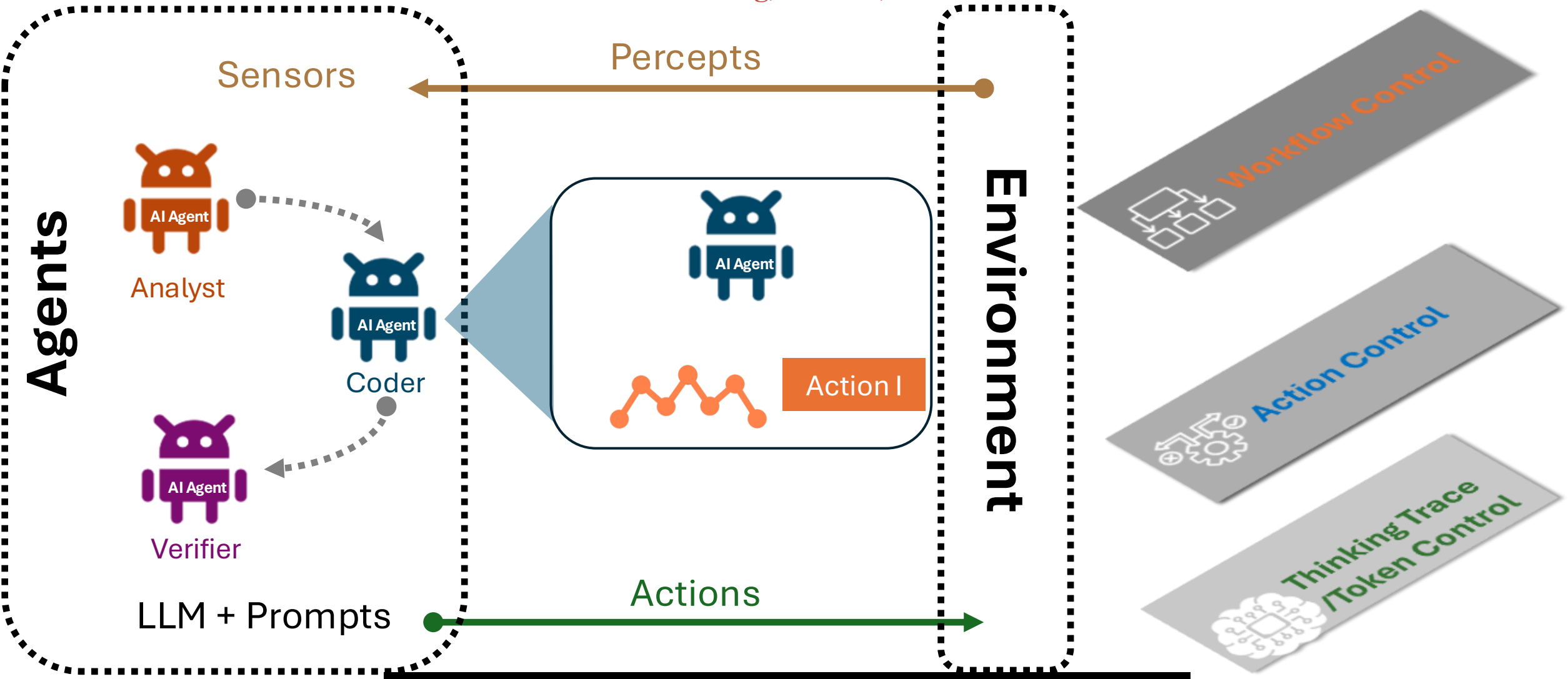
Method	MATH		AMC		MBPP		DROP		MMLU Pro		Average	
	Perf.	Cost	Perf.	Cost	Perf.	Cost	Perf.	Cost	Perf.	Cost	Perf.	Cost
IO	56.58	0.51	53.59	0.53	72.92	0.08	75.49	0.08	50.89	0.03	60.44	0.22
Manually Designed Agentic Workflows												
CoT	55.76	0.53	52.47	0.54	74.29	0.08	77.96	0.11	51.35	0.04	60.98	0.23
ComplexCoT	54.39	0.62	49.52	0.64	71.68	0.10	77.74	0.33	64.03	0.35	63.85	0.42
Self-Consistency	57.82	3.38	53.89	3.44	73.98	0.52	78.33	0.74	50.93	0.34	61.49	1.51
Multi-agent Debate	58.92	7.66	53.82	7.96	74.68	0.86	78.65	1.64	57.56	0.88	63.86	3.45
Self-Refine	55.14	1.08	52.88	1.06	71.07	0.27	76.71	0.34	43.75	0.46	57.96	0.62
MedPrompt	56.69	5.64	55.19	5.49	71.85	0.61	81.77	1.75	52.79	0.63	62.77	2.58
MultiPersona	57.41	0.56	53.74	0.58	70.38	0.28	76.13	0.29	61.70	0.36	63.87	0.41
Automated Agentic Workflow Optimization Methods												
GPTSwarm	54.94	1.69	55.42	9.04	71.26	**0.25**	78.37	**0.51**	58.37	1.14	63.43	2.54
ADAS	56.17	2.95	47.33	3.22	72.34	0.77	81.17	0.98	59.71	1.05	63.22	1.71
AgentSquare	62.76	**1.35**	58.78	**1.79**	72.73	0.52	76.84	0.76	63.75	**0.75**	66.70	**1.02**
AFlow (Qwen3-8B)	63.24	3.23	62.75	2.51	79.37	1.02	77.08	1.05	64.62	1.47	68.56	1.79
AFlow (GPT-4o)	63.99	2.77	63.77	4.86	83.77	0.57	80.26	1.02	65.61	0.89	70.40	1.95
MaAS	65.02	2.43	63.36	2.70	79.08	2.04	76.64	1.64	62.31	1.29	68.09	1.90
ScoreFlow	62.00	2.36	60.15	3.51	83.63	1.70	82.10	1.42	64.58	2.62	69.51	2.37
FlowBank	**69.34**	1.78	**67.94**	2.49	**84.26**	1.62	**83.49**	1.06	**67.40**	1.52	**73.40**	1.65



YDYCRH, FlowBank: Query-Adaptive Agentic Workflows Optimization through Precompute-and-Reuse, arXiv:2606.11290.

Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow



Precompute and Reuse

Toward Self-Improving Agentic Systems

Across Thinking, Actions, and Workflow

1. **[Transfer Q*]** Chakraborty, Ghosal, Yin, Manocha, Wang, Bedi, and Huang. "Transfer Q-star: Principled Decoding for LLM Alignment." NeurIPS 2024.
2. **[GenARM]** Xu, Sehwag, Koppel, Zhu, An, Huang, and Ganesh. "GenARM: Reward Guided Generation with Autoregressive Reward Model for Test-time Alignment." ICLR 2025.
3. **[ACT]** Liu, Liu, Ho, Chakraborty, Wang, and Huang. "Agentic Critical Training." arXiv:2603.08706.
4. **[FlowBank]** Yuan, Deng, Yu, Chakraborty, Rostami, and Huang. "FlowBank: Query-Adaptive Agentic Workflows Optimization through Precompute-and-Reuse." arXiv:2606.11290.
5. **[AegisLLM]** Cai, Shabihi, An, Che, Bartoldson, Kailkhura, Goldstein, Huang. "AegisLLM: Scaling Agentic Systems for Self-Reflective Defense in LLM Security." ICLR workshop 2025.

furongh@umd.edu https://furong-huang.com/ X: @furongh