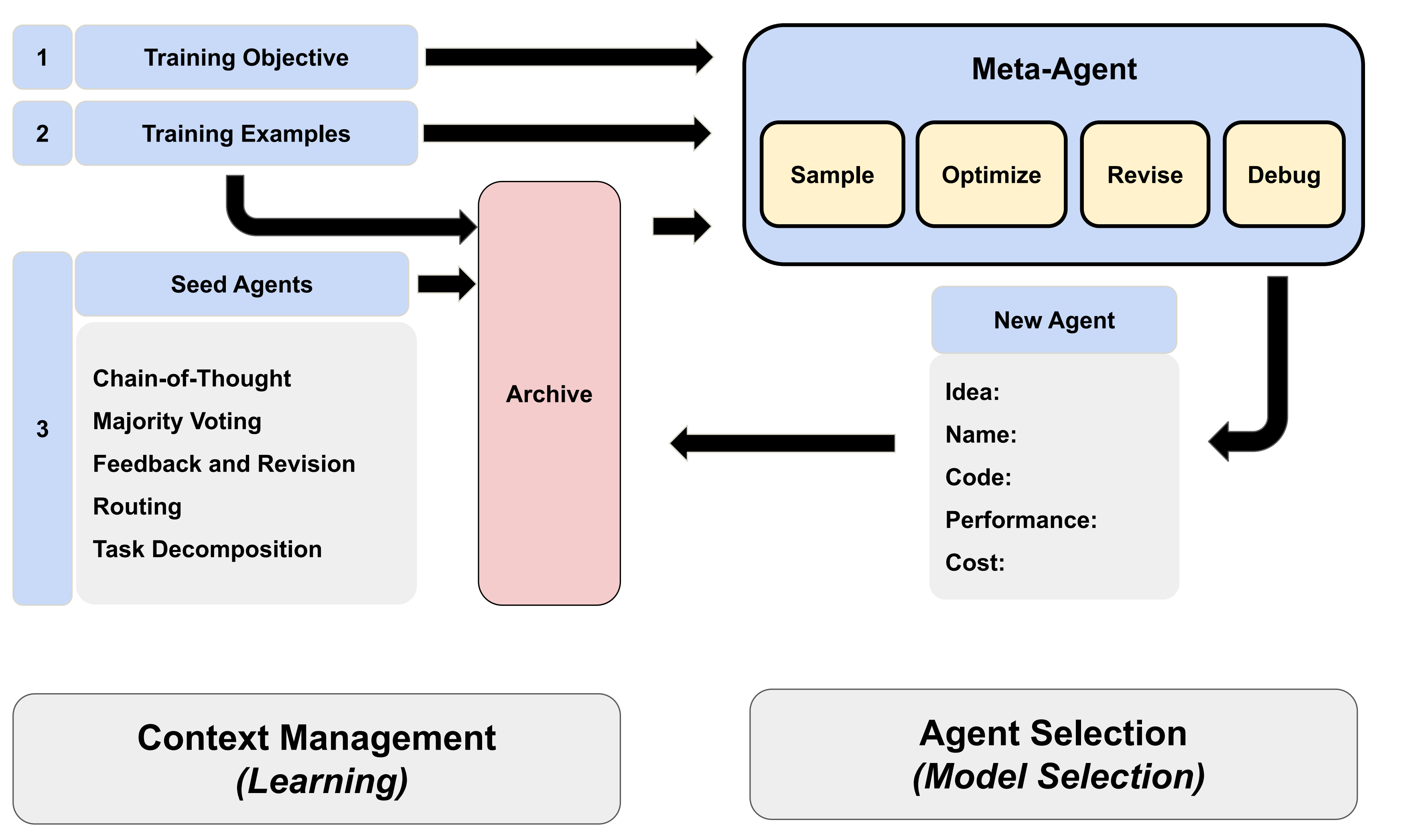
Exploring Two Open Questions in Meta-Agent Design

Overview



1 Training Objective

2 Training Examples

3 Seed Agents

Chain-of-Thought
Majority Voting
Feedback and Revision
Routing
Task Decomposition

Archive

Meta-Agent

Sample | Optimize | Revise | Debug

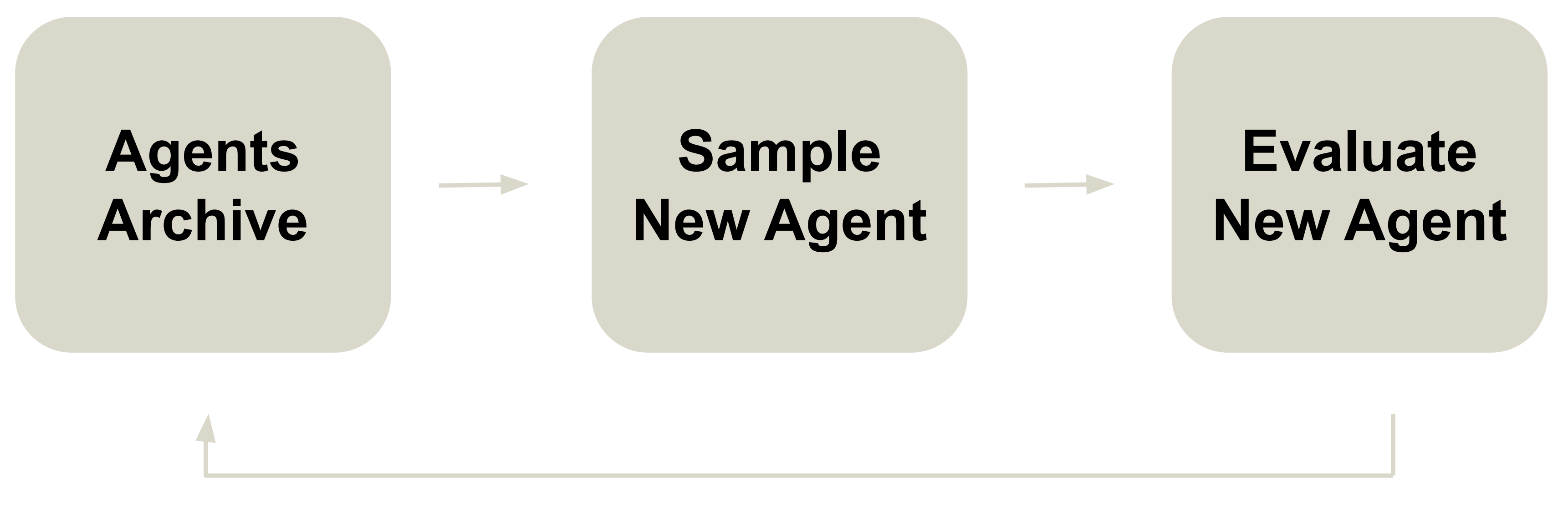
New Agent

Idea:
Name:
Code:
Performance:
Cost:

Context Management
(Learning)

Agent Selection
(Model Selection)

Background



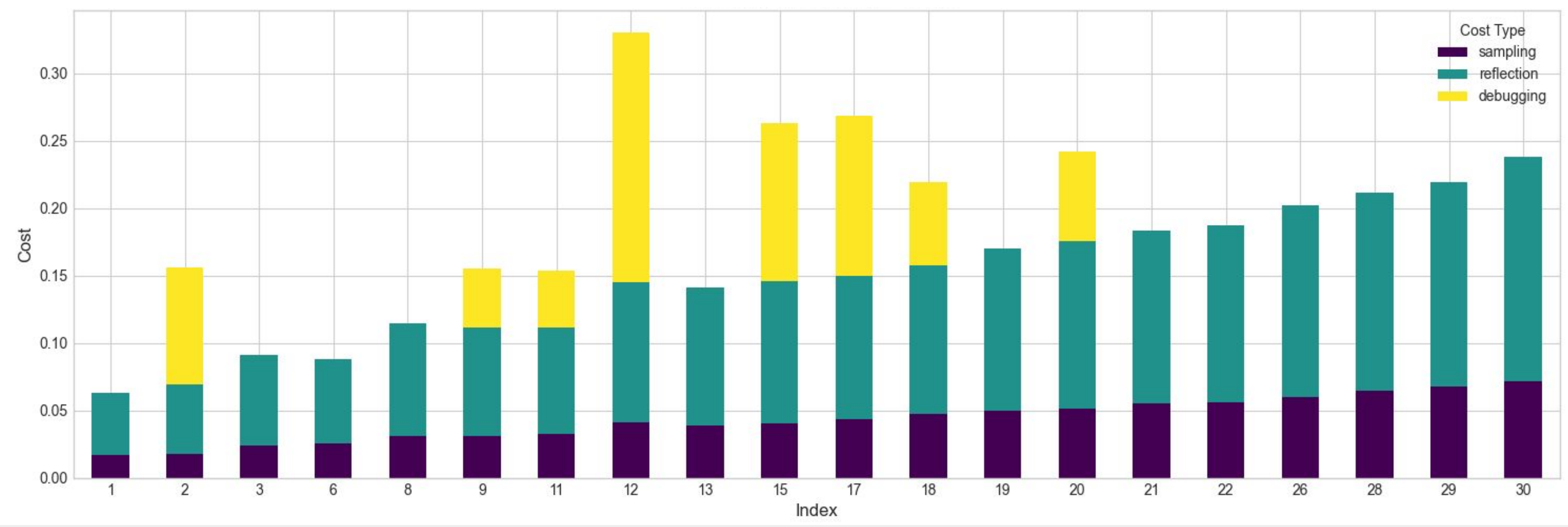
Agents Archive → Sample New Agent → Evaluate New Agent

- Automated Design of Agentic Systems
- Archon: An Architecture Search Framework for Inference-Time Techniques
- Self-Taught Optimizer (STOP): Recursively Self-Improving Code Generation
- Large Language Models as Tool Makers
- AFlow: Automating Agentic Workflow Generation

Research Problems

1 How can we mitigate ***error accumulation*** and ***cost accumulation*** while allowing the meta-agent to learn from previous attempts? → **DCM**

2 How can we better tackle the ***agent selection*** problem, which is analogous to the model selection problem in machine learning? → **RADAR**



Simple Abstractions (SA)

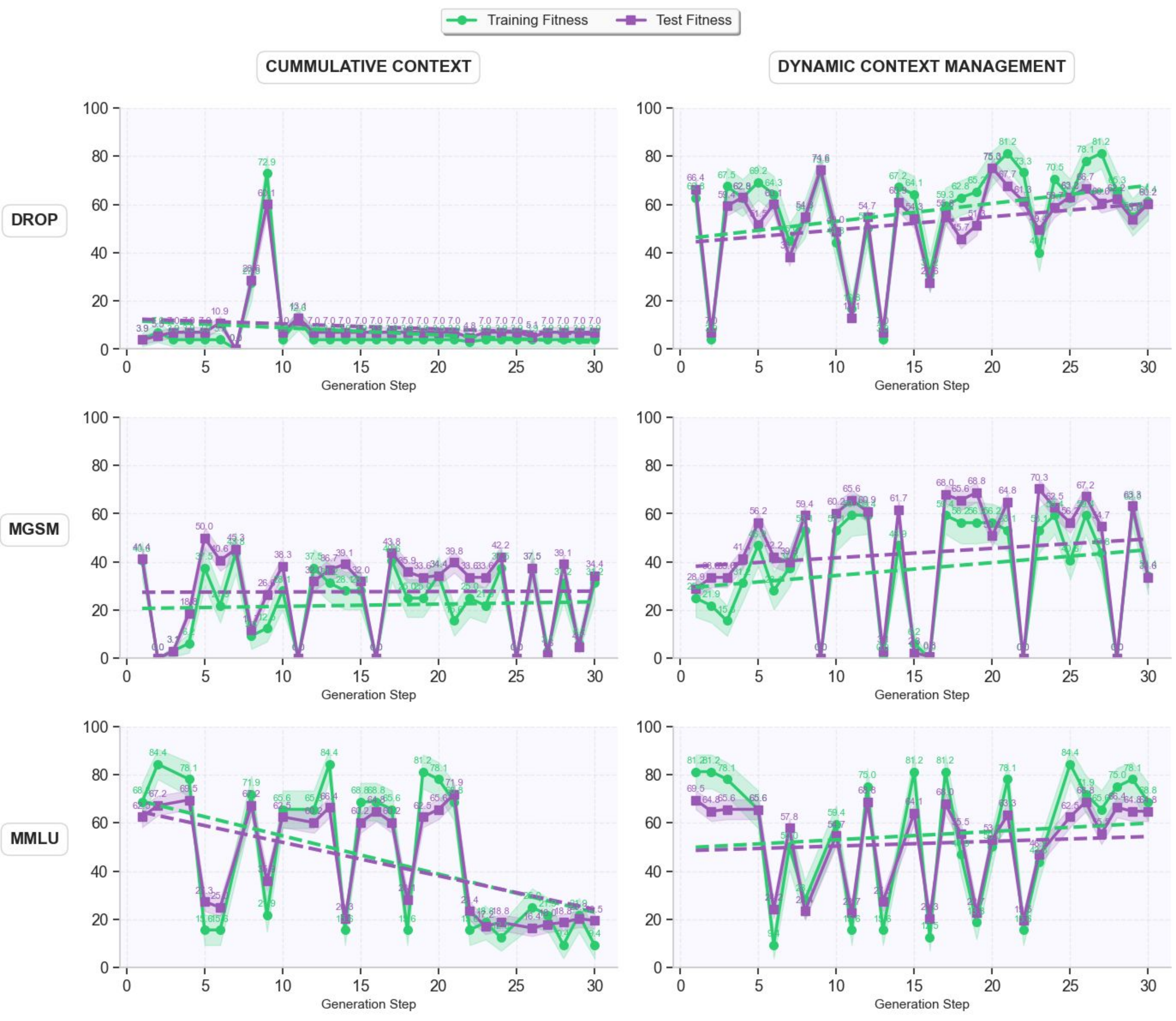
```python
import LanguageModel
output_fields = ["reasoning", "answer"] # Example output fields
lm_agent = LanguageModel(output_fields, mode="standard")
instruction = "Please think step by step and then solve the task."
task_context = "Solve the equation x^2 - 4 = 0 for real x."
# Directly unpack the fields using the callable interface
reasoning, answer = lm_agent(task_context, instruction)

class AgentSystem:
    def __init__(self):
        # Initialize LanguageModel instances here.
        pass
    def forward(self, prompt: str):
        # Abstract method to be implemented by subclasses.
        # Args: prompt (str): The input prompt for the agent.
        # Returns: str: The agent's response.
        raise NotImplementedError("Subclasses must implement.")
```

Dataset	ADAS Abstractions (172 lines)		Simple Abstractions (17 lines)	
	Agent Name	Test Acc.	Agent Name	Test Acc.
MGSM	LLM Debate	47.7 ± 4.3	LLM Debate	55.47 ± 4.3
MMLU	LLM Debate	72.7 ± 3.9	Chain of Thought	75.0 ± 3.7
DROP	Self-Quality-Diversity	67.2 ± 1.2	Majority Vote	73.93 ± 3.6

Table 2: Performance of Initial Agents with Simpler Abstractions, Test Acc. ± Std

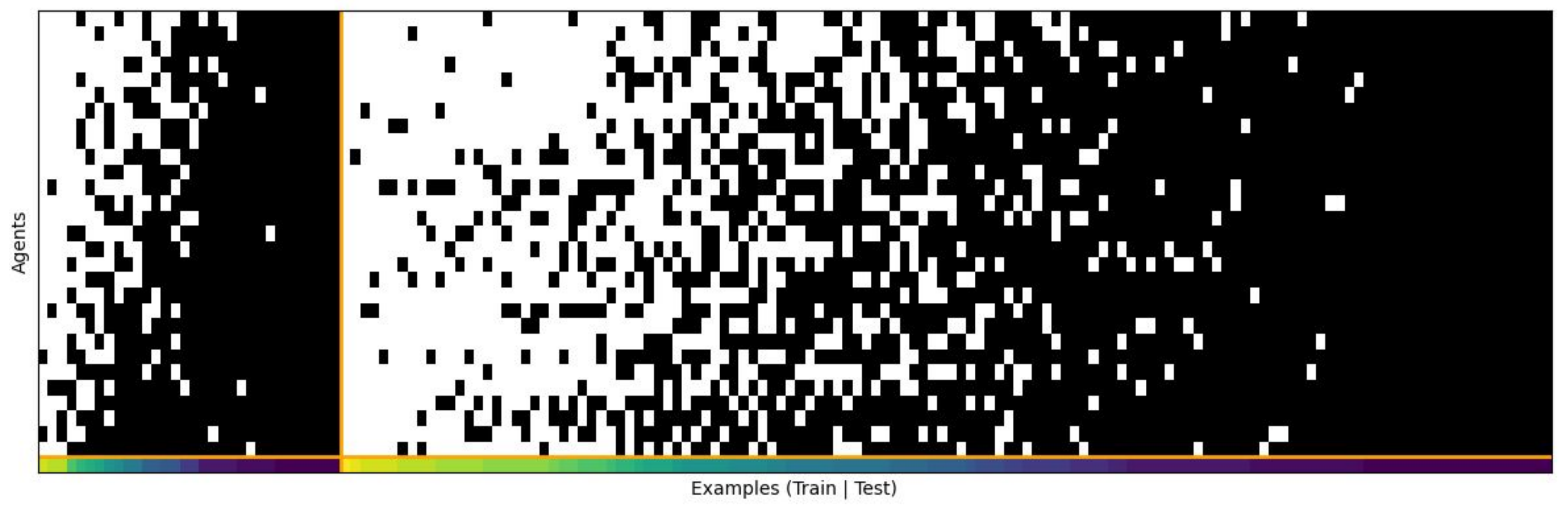
Dynamic Context Management (DCM)

Dataset	CoT	ADAS	DCM	SA+DCMS
MGSM	35.9 ± 4.1	57.0 ± 4.3	66.4 ± 4.1	$\mathbf{70.3} \pm 3.9$
MMLU	64.8 ± 4.1	74.2 ± 3.7	$\mathbf{75.0} \pm 3.7$	$\mathbf{75.0} \pm 3.7$
DROP	63.5 ± 1.0	67.2 ± 1.2	74.0 ± 1.2	$\mathbf{75.3} \pm 3.5$

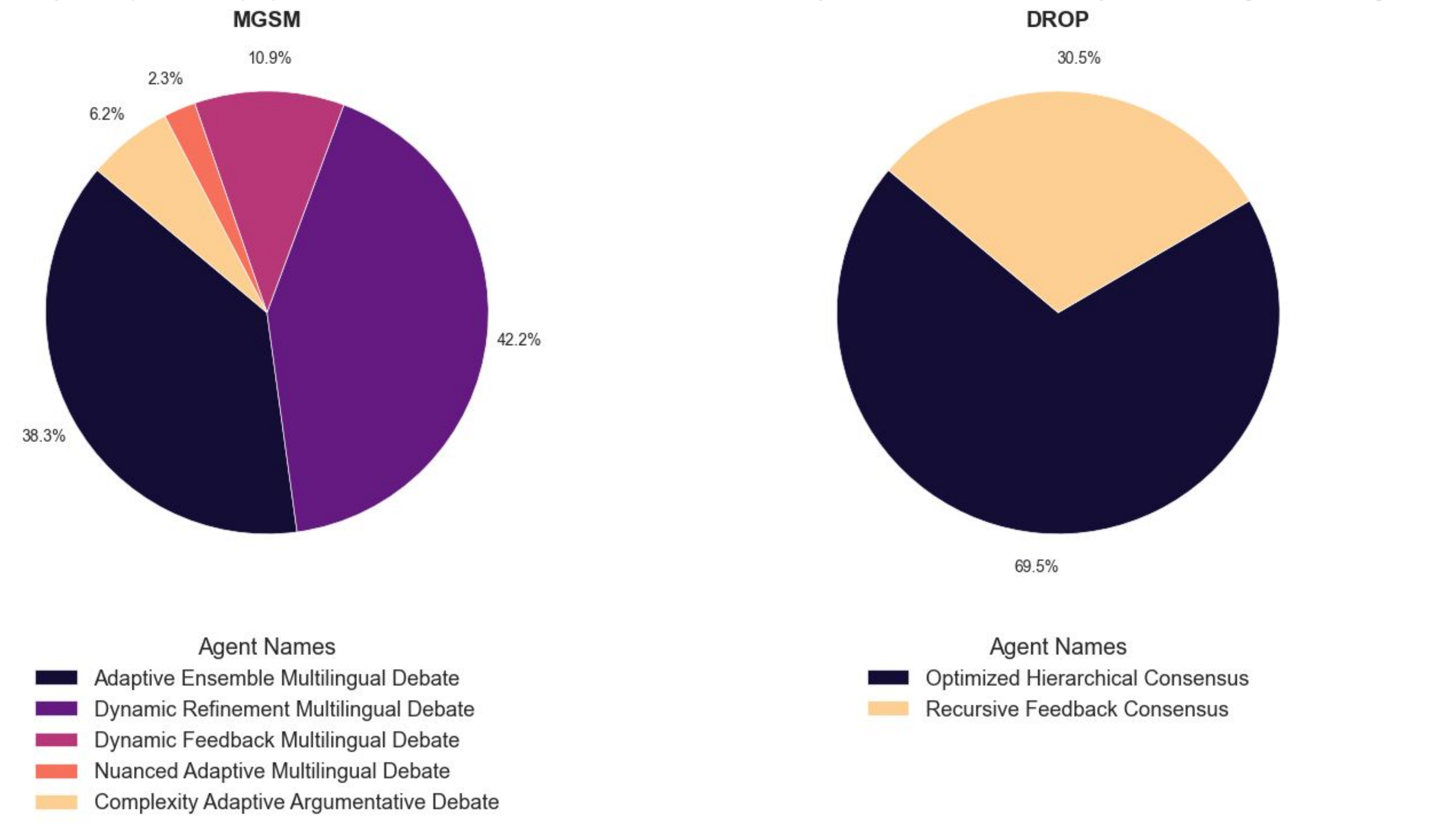


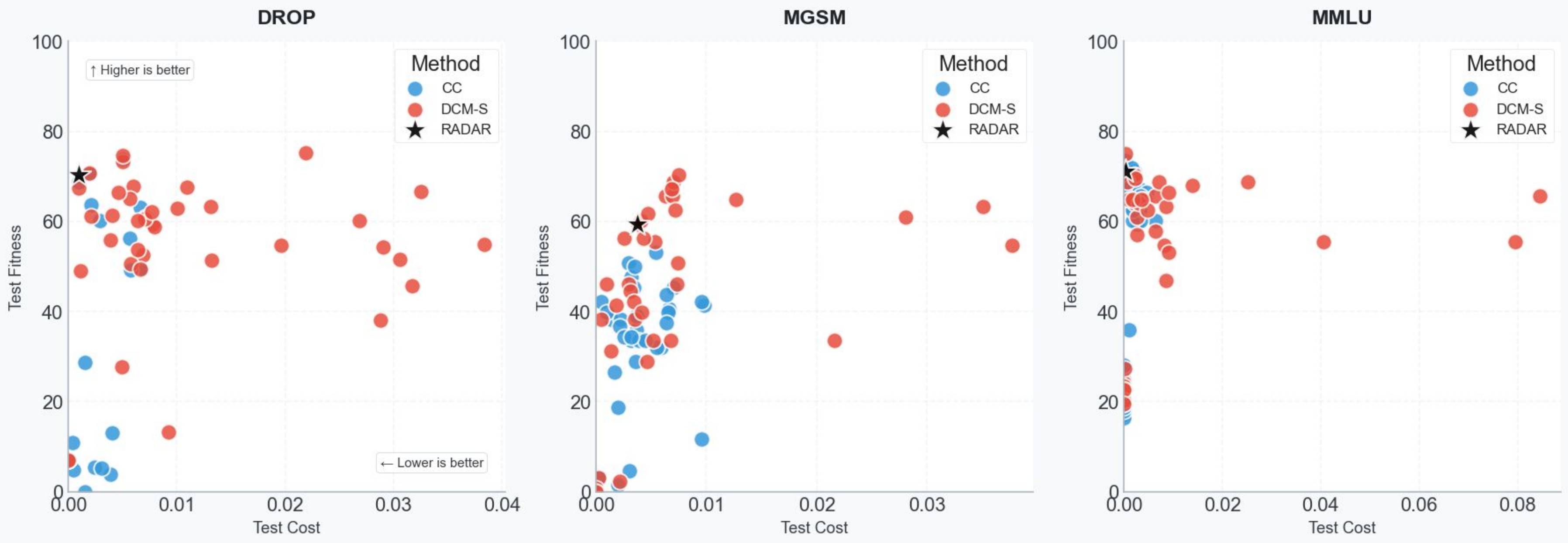
Dataset	Baseline	Selected on Training			Selected on Test		
	CoT	Initial	ADAS	DCM	Initial	ADAS	DCM
MGSM	35.9 ± 4.1	47.7 ± 4.3	48.4 ± 4.3	$\mathbf{64.8} \pm 4.1$	47.7 ± 4.3	57.0 ± 4.3	$\mathbf{66.4} \pm 4.1$
MMLU	64.8 ± 4.1	65.6 ± 4.1	65.6 ± 4.1	65.6 ± 4.1	72.7 ± 3.9	74.2 ± 3.7	$\mathbf{75.0} \pm 3.7$
DROP	63.5 ± 1.0	67.2 ± 1.2	64.5 ± 1.0	$\mathbf{74.0} \pm 1.2$	67.2 ± 1.2	67.2 ± 1.2	$\mathbf{74.0} \pm 1.2$

Retrieval Augmented Routing (RADAR)



Dataset	CoT	ADAS	SA+DCMS	SA+DCMS+R
MGSM	35.9 ± 4.1	48.4 ± 4.3	63.28 ± 4.1	$\mathbf{71.09} \pm 3.9$
DROP	63.5 ± 1.0	64.5 ± 1.0	67.68 ± 3.85	64.57 ± 3.9


MGSM
DROP
Agent Names
Adaptive Ensemble Multilingual Debate
Dynamic Refinement Multilingual Debate
Dynamic Feedback Multilingual Debate
Nuanced Adaptive Multilingual Debate
Complexity Adaptive Argumentative Debate
Agent Names
Optimized Hierarchical Consensus
Recursive Feedback Consensus


DROP
MGSM
MMLU
↑ Higher is better
← Lower is better
Method
CC
DCM-S
RADAR
Test Fitness
Test Cost