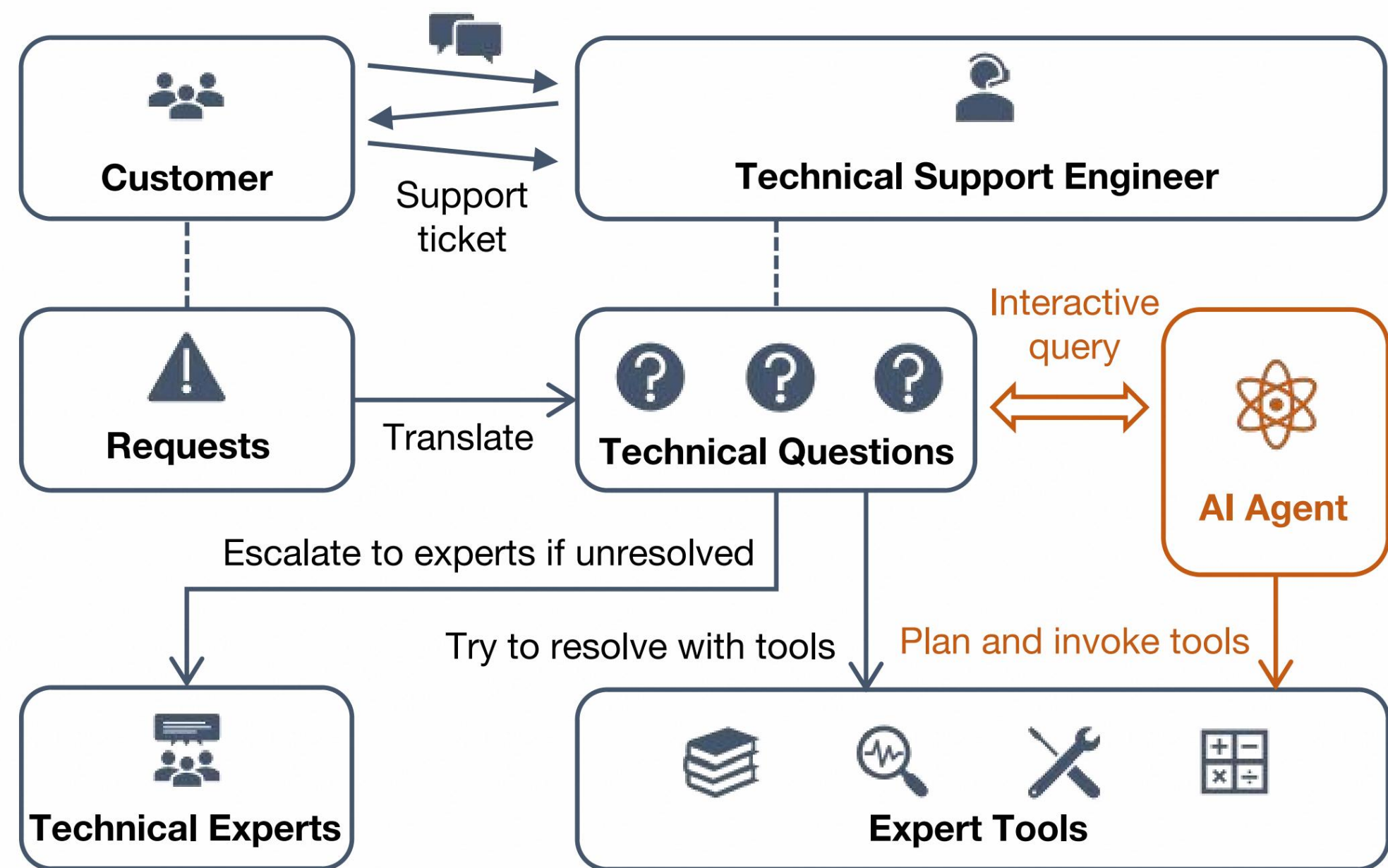


Haoyu Wang, Wenxuan Ma, Bing Hu, Haozhe Li, Qingqian Si, Hongke Guo, Boyang Huang, Zhaoliang Zhu, You Zhang, Yu Zhou, Xudong Zheng
Alibaba Cloud Computing

BACKGROUND

AIOps Agent in Ticket Support

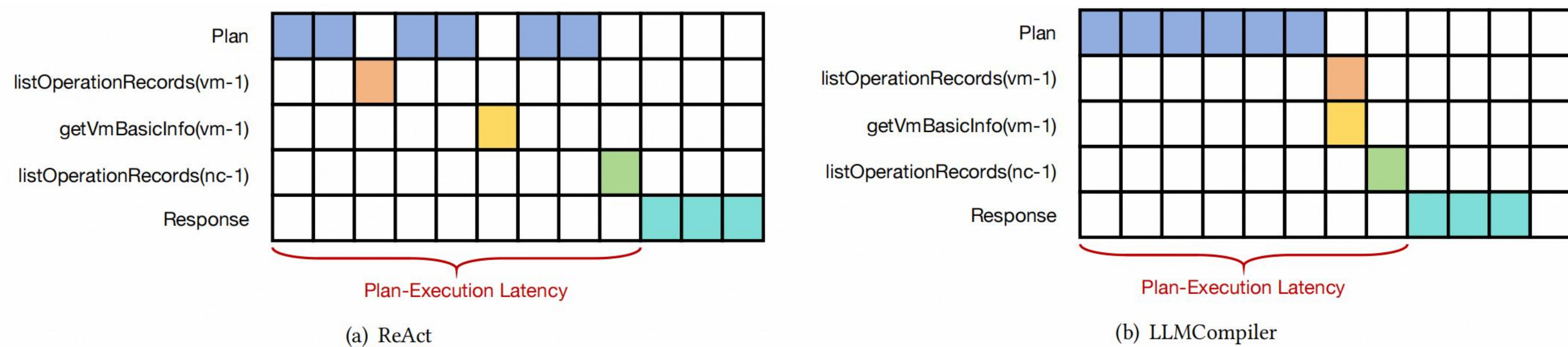
- Ticket support engineers use diverse expert tools to resolve customer queries, which can be enhanced via LLM.



MOTIVATION

Challenge I: Plan-Execution Latency

- DAG-based tool execution is optimal.
- The optimization efforts are focused on the **planning** phase.



Challenge II: Planning Inaccuracy

- Errors from **planning** are the primary reasons for inaccuracy.

Reason	ReAct	LLMCompiler
Failure in tool selection	68.3%	62.1%
Incorrect tool parameters	18.3%	18.2%
Broken inter-step dependency	3.3%	10.6%
Incorrect or incomplete conclusion	10.0%	9.1%

Opportunity: High-Freq. Scenarios

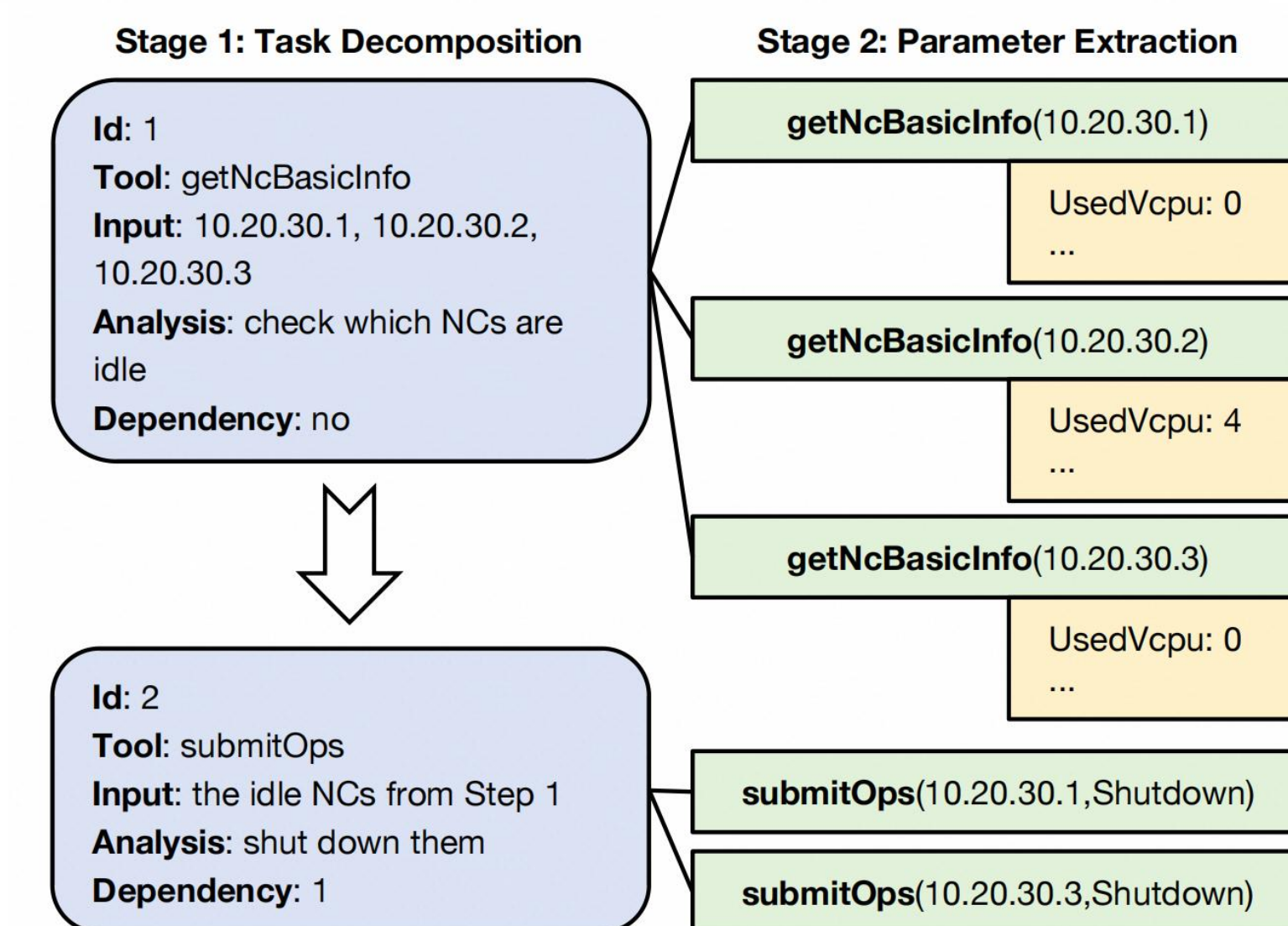
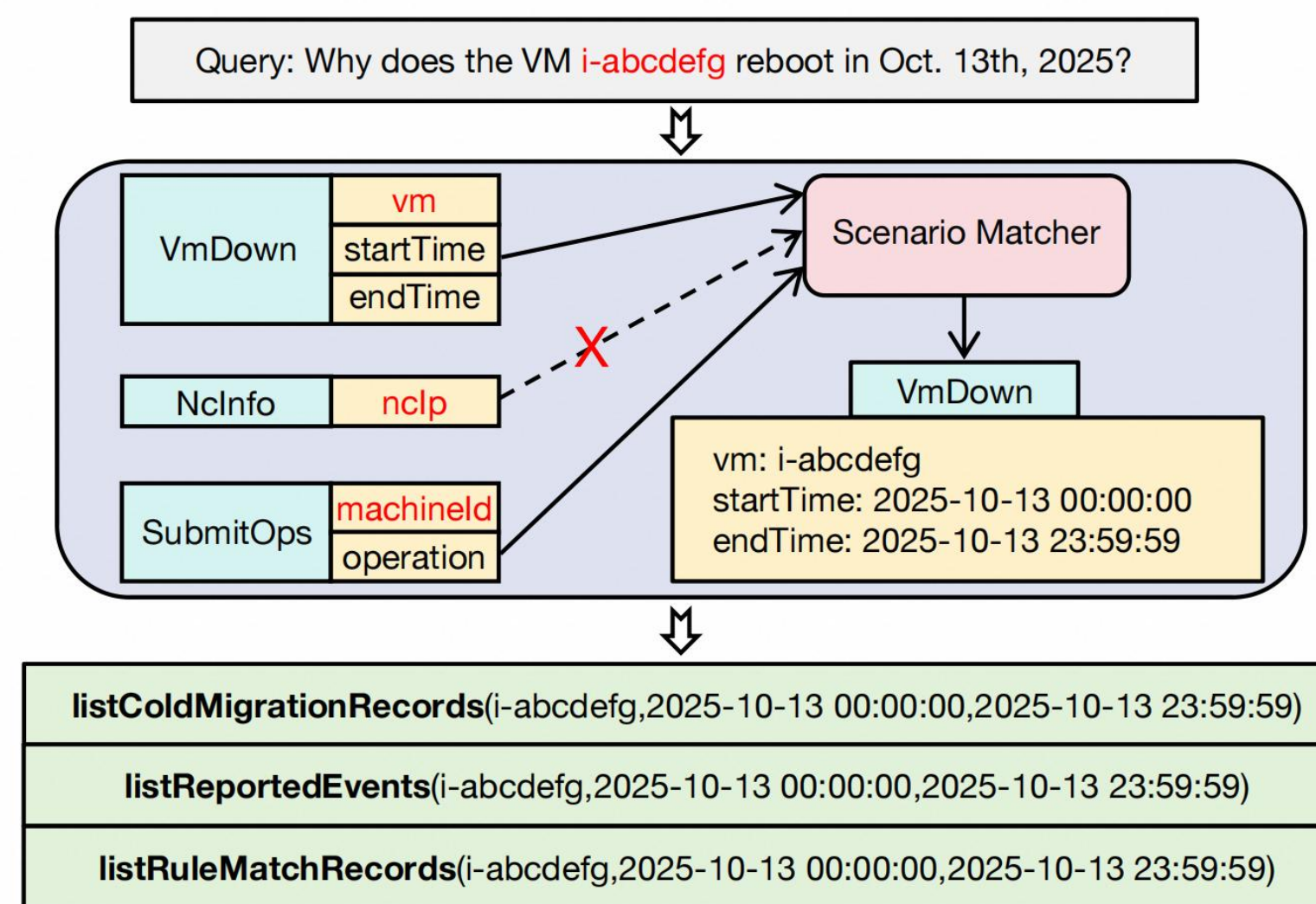
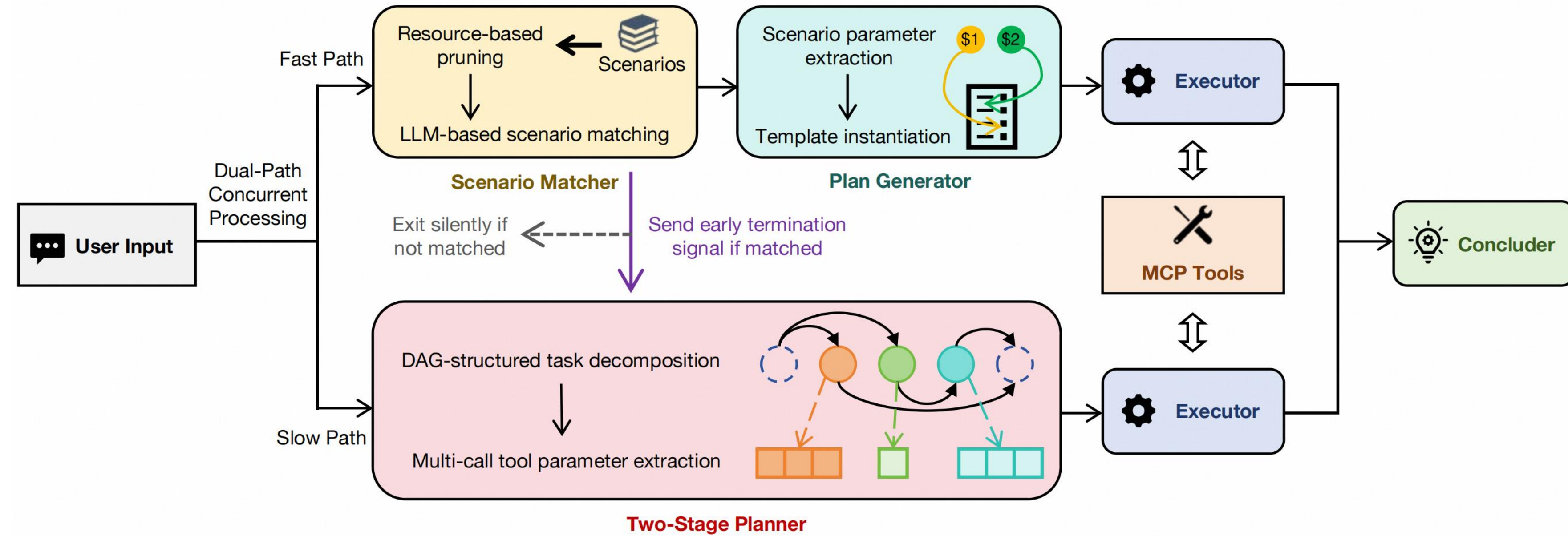
- User queries have many **high-frequency scenarios**.
- Expensive LLM-based full-plan generation is bypassed with **pre-validated templates**.

Scenario	Percentage
Query general knowledge	19.9%
Query the information of a VM	7.9%
Evaluate the success likelihood of migration	7.2%
Query the system events of a VM	6.5%
Query the information of a physical machine	5.4%

METHODS

DualLane: Dual-Path Planning Architecture

- A slow path designed for complex scenarios, and a fast path for frequent ones.



EXPERIMENTS

Offline Evaluation on Collected Dataset

- Compared to existing approaches, DualLane achieves much **improvements in both accuracy and latency**.

Online Evaluation in 30-day Production Test

- Plan-execution latency: median 4.2s, 90th percentile 10.1s
- Error rate: 26.1% total, only 7.1% attributed to DualLane

Difficulty	Method	Latency		Acc.
		Total Resp.	Plan-Exec.	
Easy	DualLane	6.89s	3.69s	97.6%
	LLMCompiler	13.15s	4.34s	84.6%
	ReAct	7.64s	4.33s	85.6%
Medium	DualLane	13.69s	7.08s	97.5%
	LLMCompiler	20.43s	9.64s	69.0%
	ReAct	18.36s	14.23s	71.4%
Hard	DualLane	13.06s	7.29s	92.3%
	LLMCompiler	14.28s	5.88s	64.1%
	ReAct	14.77s	8.92s	83.3%
All	DualLane	9.73s	5.22s	96.5%
	LLMCompiler	15.04s	5.87s	76.7%
	ReAct	11.58s	7.55s	81.9%