FlowXpert: Expertizing Troubleshooting Workflow Orchestration with Knowledge Base and Multi-Agent Coevolution

Binpeng Shi[1], Yu Luo[1], Jingya Wang[1], Yongxin Zhao[1], Shenglin Zhang[1], Bowen Hao[1], Chenyu Zhao[1], Yongqian Sun[1], Zhi Zhang[2], Ronghua Sun[2], Haihua Li[2], Wei Song[2], Xiaolong Chen[2], Jingbo Miao[2], Dan Pei[3]

[1]Nankai University, [2]Huawei, [3]Tsinghua University

- Reporter: Binpeng Shi
- Email: shibinpeng23@mail.nankai.edu.cn

Outline

What are the typical resolutions for a cloud incident?

➔ Workflow, step-by-step guidance and executable scripts

How to transform a naive LLM into a workflow generator in the field of cloud services?

➔ Support of domain knowledge, alignment of application capability

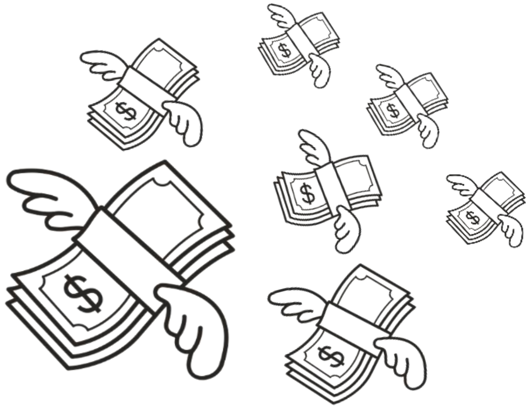
Framework design

➔ Knowledge Base Construction, Multi-Agent Coevolution

Evaluation

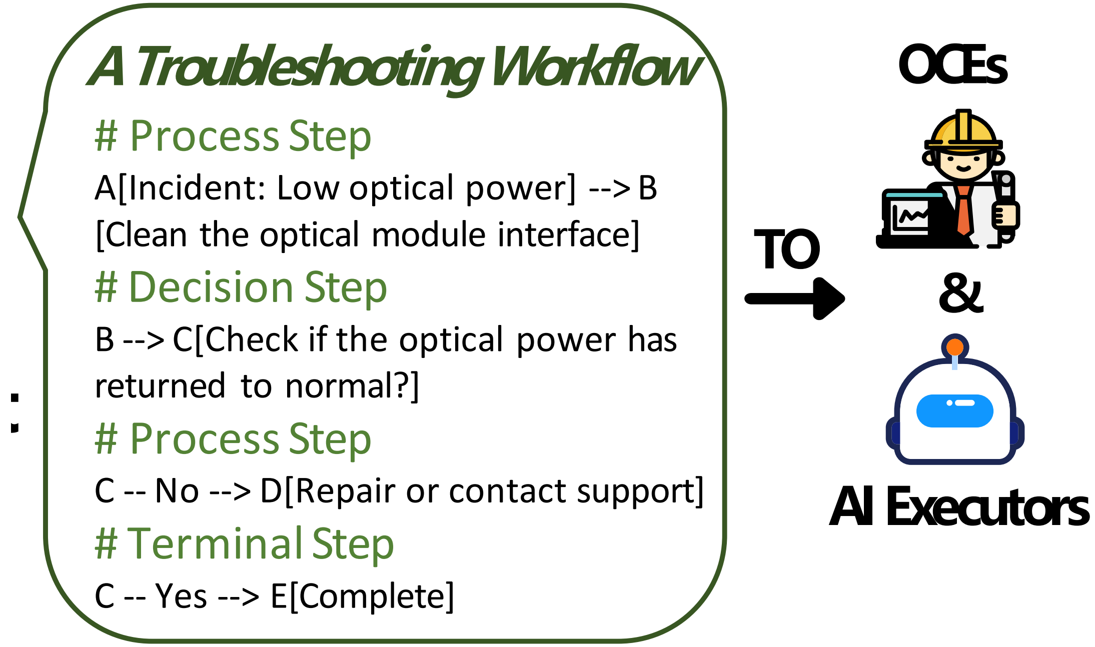
➔ Benchmark tests, online deployment, case study

Impact of Incidents



Incidents → Unsatisfying customers → Economic loss

Typical Resolutions for Cloud Incidents



A Troubleshooting Workflow

```
# Process Step
A[Incident: Low optical power] --> B
[Clean the optical module interface]
# Decision Step
B --> C[Check if the optical power has
returned to normal?]
# Process Step
C -- No --> D[Repair or contact support]
# Terminal Step
C -- Yes --> E[Complete]
```

- Step-by-step guidance

- Executable scripts
- AI agent equipped with tool invocation and result analysis capabilities

Most cloud service providers abstract troubleshooting into workflows, which follow a structured sequence of core steps

Workflow Usage and Acquisition

Heavy Usage

- Workflow Recommendations Based on Similar Cases
- Automated Incident Execution and Analysis

Difficult Acquisition

- For a workflow: 7 Hours + 7 OCEs
- Including contributions from 2 experts

Workflows play a critical role in troubleshooting, which urgently needs to shift from manual creation to automated orchestration

Outline

What are the typical resolutions for a cloud incident?

➔ Workflow, step-by-step guidance and executable scripts

How to transform a naive LLM into a workflow generator in the field of cloud services?

➔ Support of domain knowledge, alignment of application capability

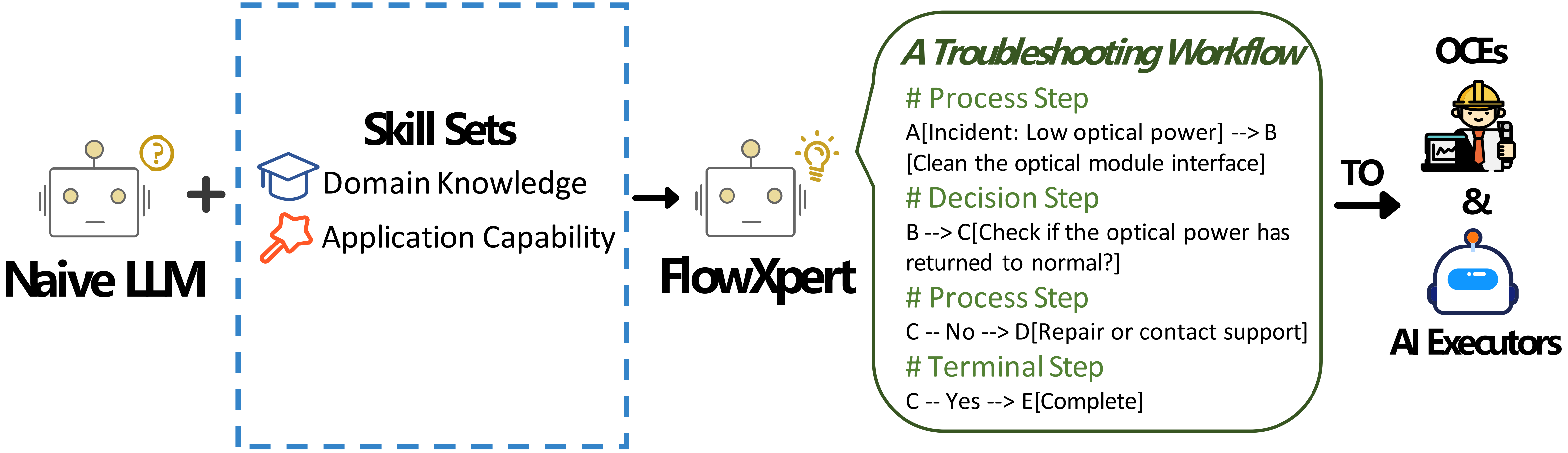
Framework design

➔ Knowledge Base Construction, Multi-Agent Coevolution

Evaluation

➔ Benchmark tests, online deployment, case study

Design Motivation



A Troubleshooting Workflow

# Process Step

A[Incident: Low optical power] --> B
[Clean the optical module interface]

# Decision Step

B --> C[Check if the optical power has returned to normal?]

# Process Step

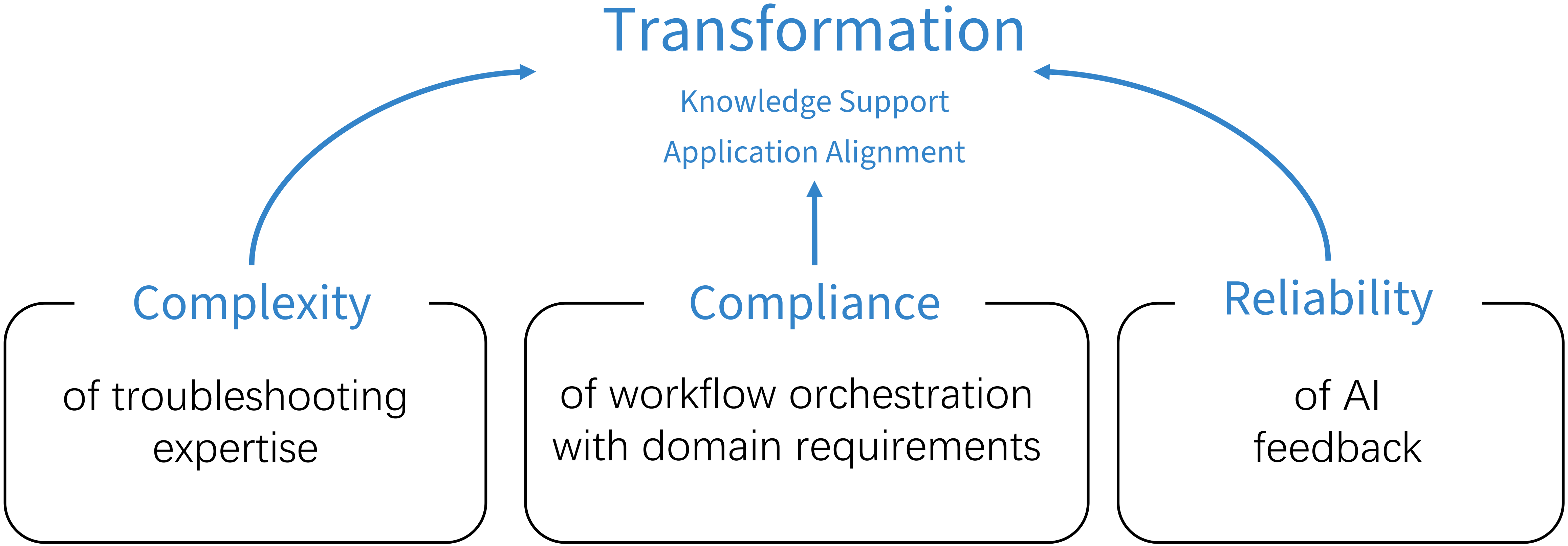
C -- No --> D[Repair or contact support]

# Terminal Step

C -- Yes --> E[Complete]

From naive LLM to workflow generator

Challenges in the Transformation Process



Transformation

Knowledge Support

Application Alignment

Complexity

of troubleshooting expertise

Compliance

of workflow orchestration with domain requirements

Reliability

of AI feedback

Outline

What are the typical resolutions for a cloud incident?

➔ Workflow, step-by-step guidance and executable scripts

How to transform a naive LLM into a workflow generator in the field of cloud services?

➔ Support of domain knowledge, alignment of application capability

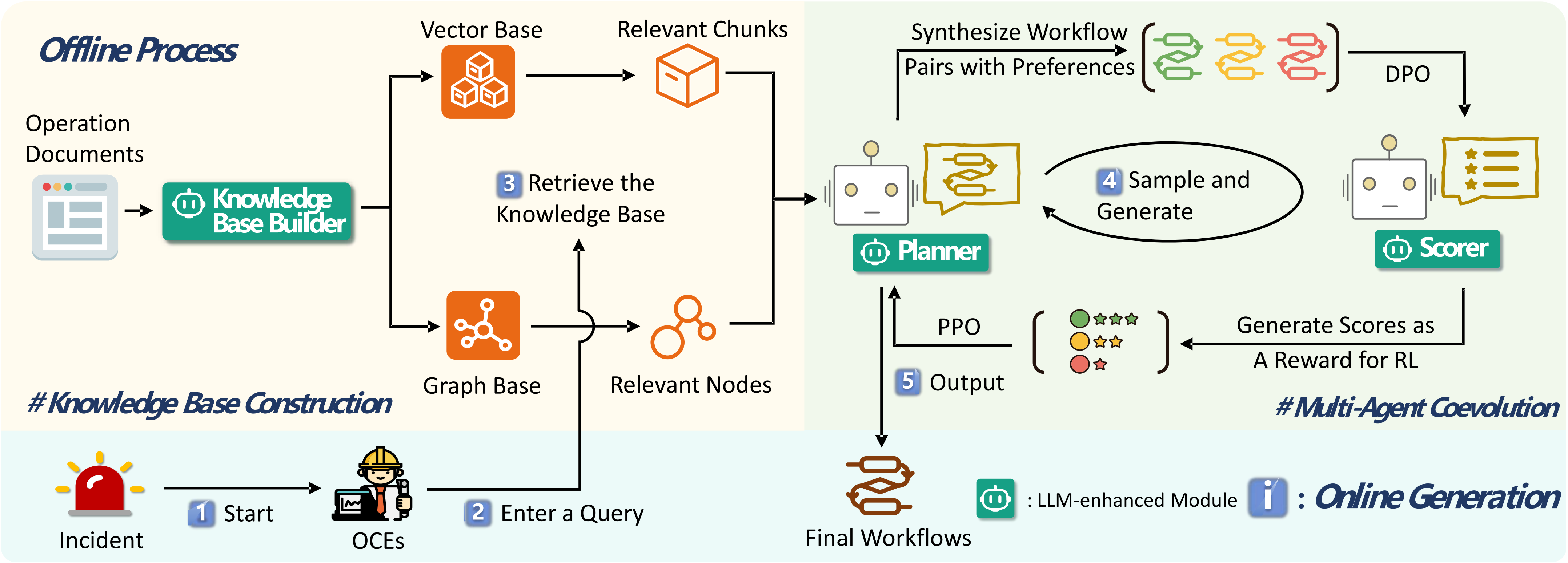
Framework design

➔ Knowledge Base Construction, Multi-Agent Coevolution

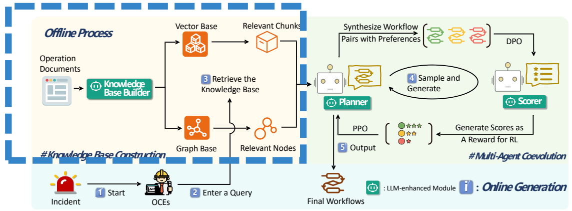
Evaluation

➔ Benchmark tests, online deployment, case study

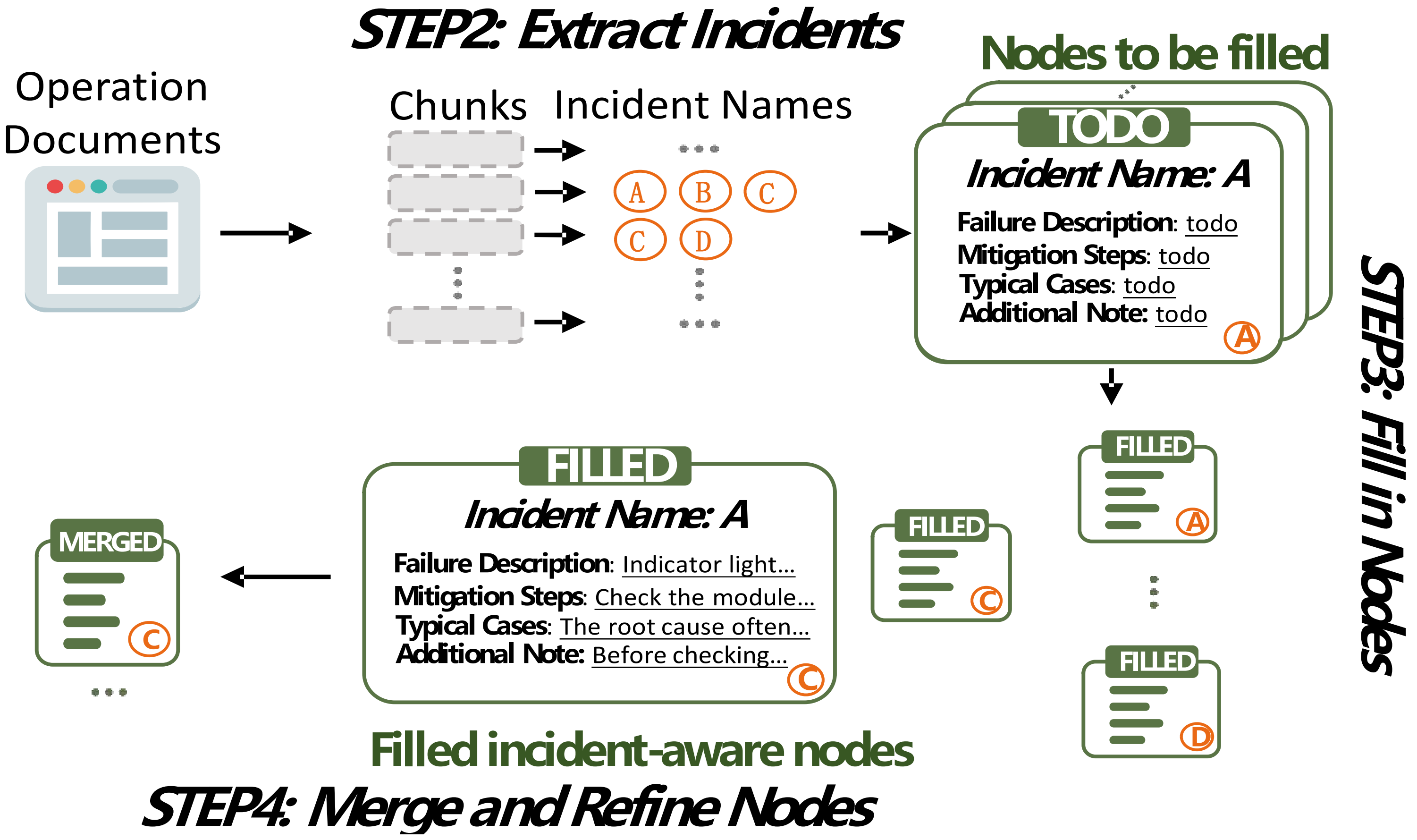
FlowXpert Overview



A Framework dedicated to transforming naive LLMs into high-quality workflow generators



FlowXpert - Module#1



TODO

Incident Name: A

Failure Description: todo
Mitigation Steps: todo
Typical Cases: todo
Additional Note: todo

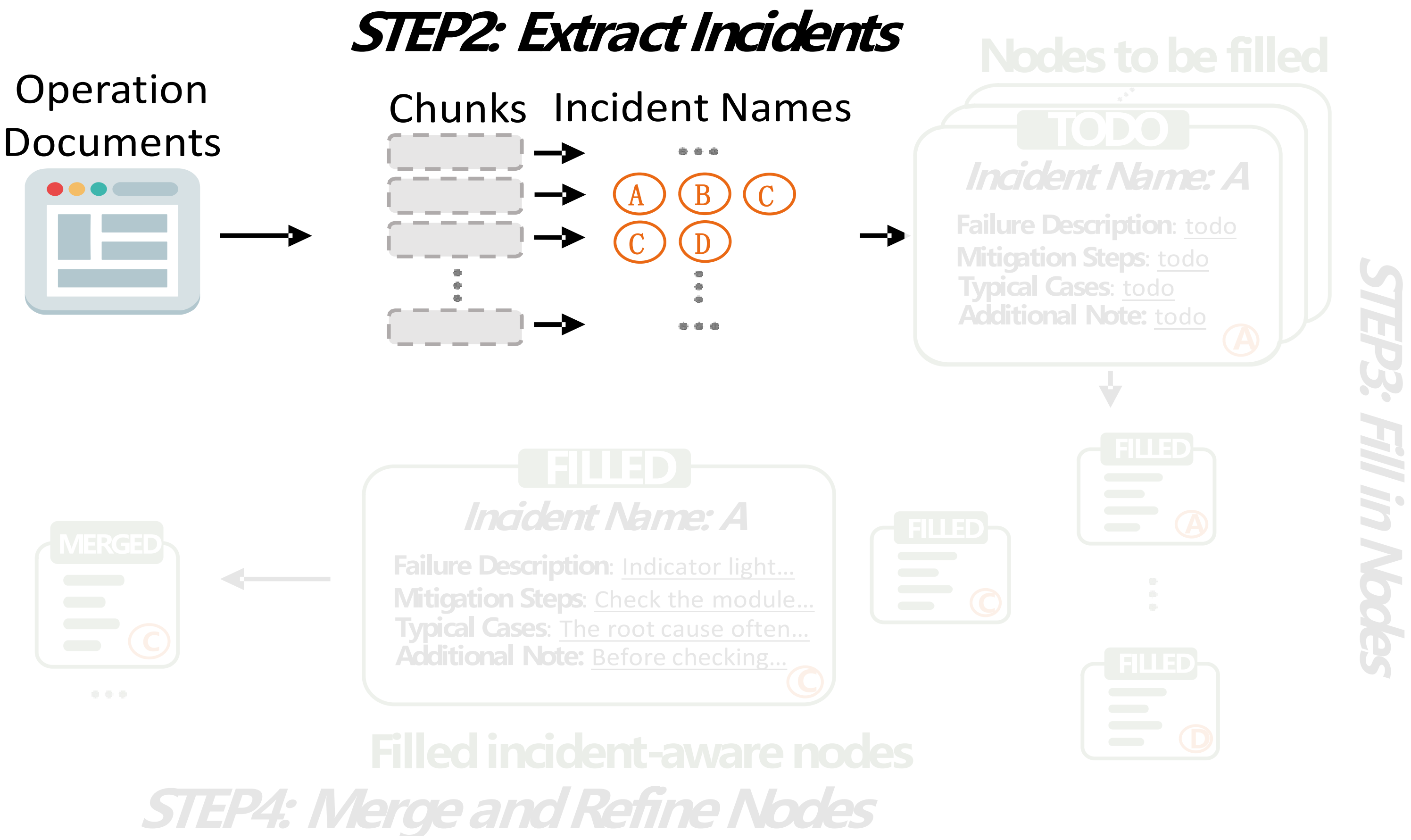
FILLED

Incident Name: A

Failure Description: Indicator light...
Mitigation Steps: Check the module...
Typical Cases: The root cause often...
Additional Note: Before checking...

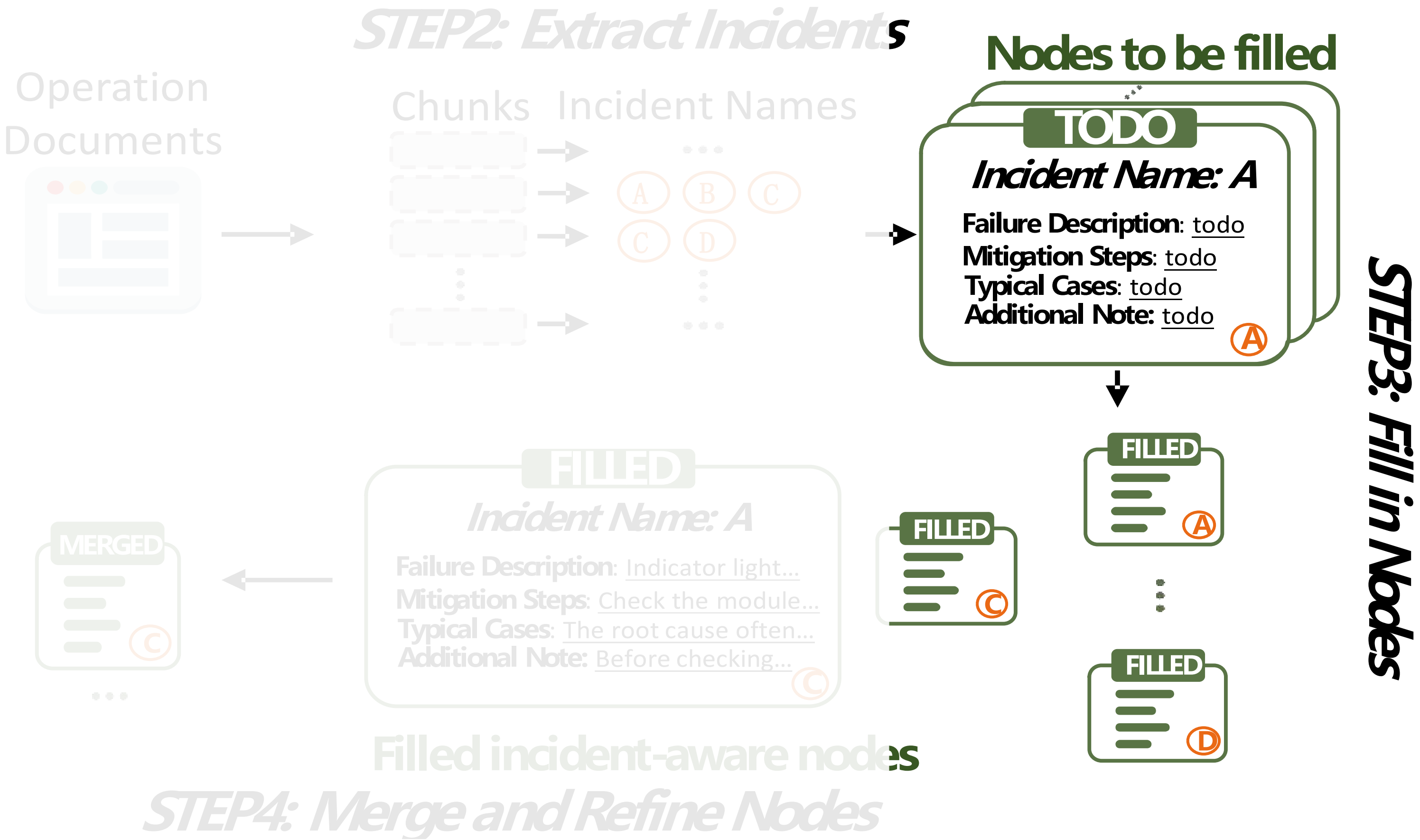
Module #1 Knowledge Base Construction

FlowXpert - Module#1



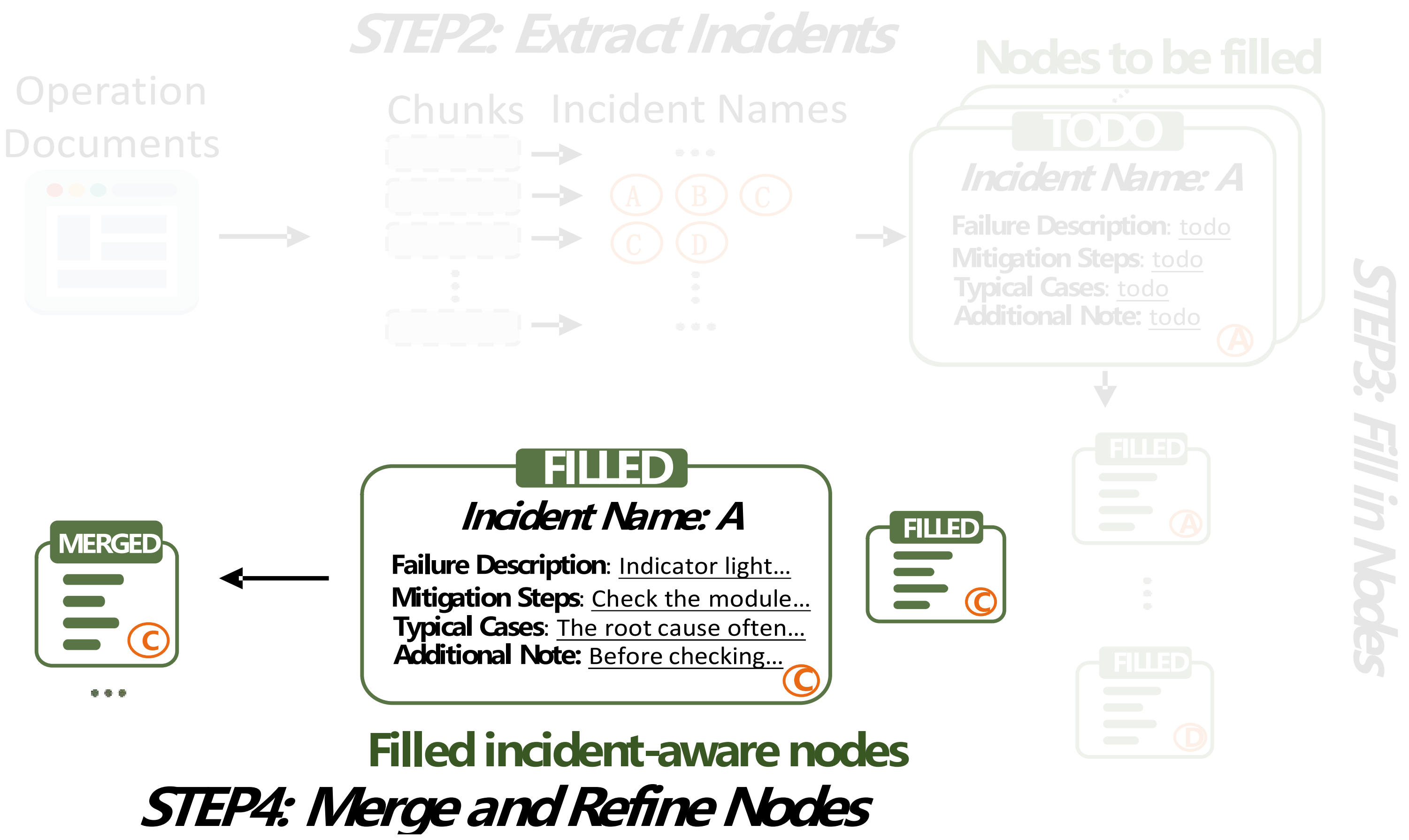
Module #1 Extract Incidents from Chunks

FlowXpert - Module#1

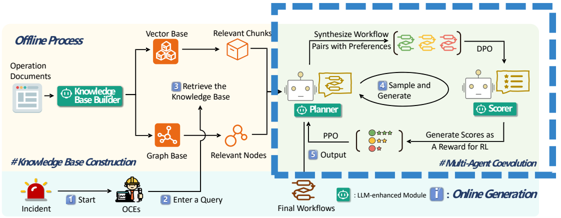


Module #1 Fill in Incident-Aware Nodes

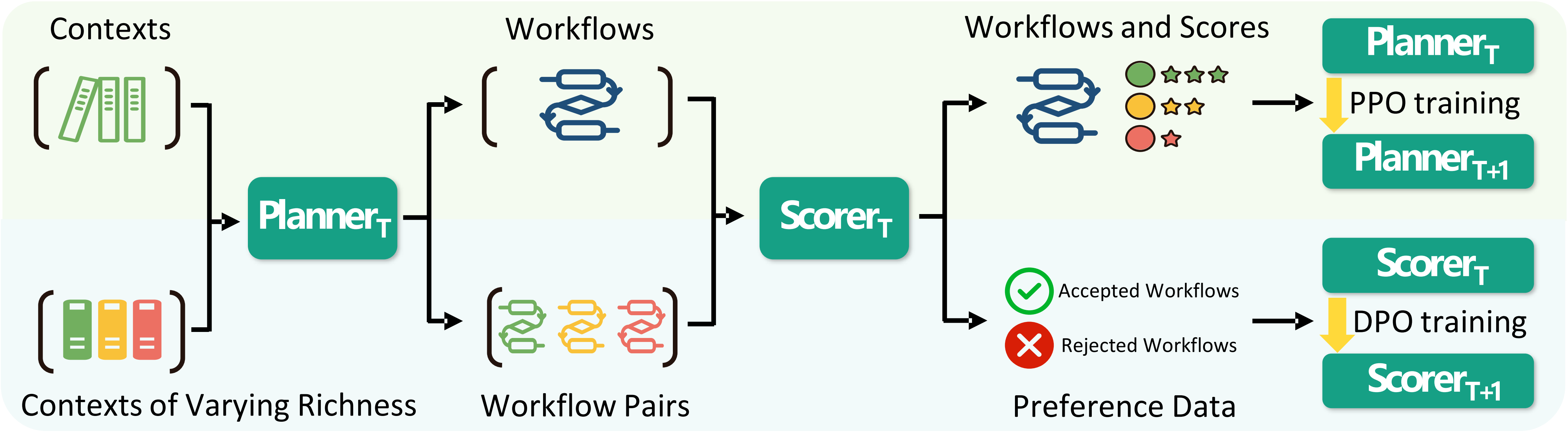
FlowXpert - Module#1



Module #1 Merge and Refine Nodes across Chunks



FlowXpert - Module#2



Contexts

Contexts of Varying Richness

Planner_T

Workflows

Workflow Pairs

Scorer_T

Workflows and Scores

Accepted Workflows

Rejected Workflows

Preference Data

Planner_T

PPO training

Planner_{T+1}

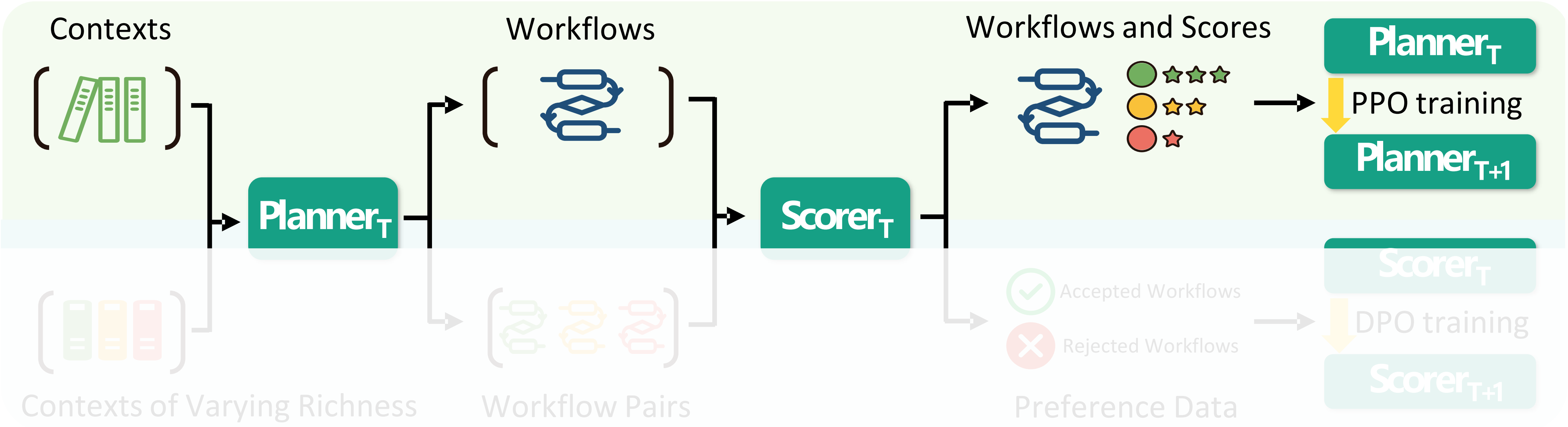
Scorer_T

DPO training

Scorer_{T+1}

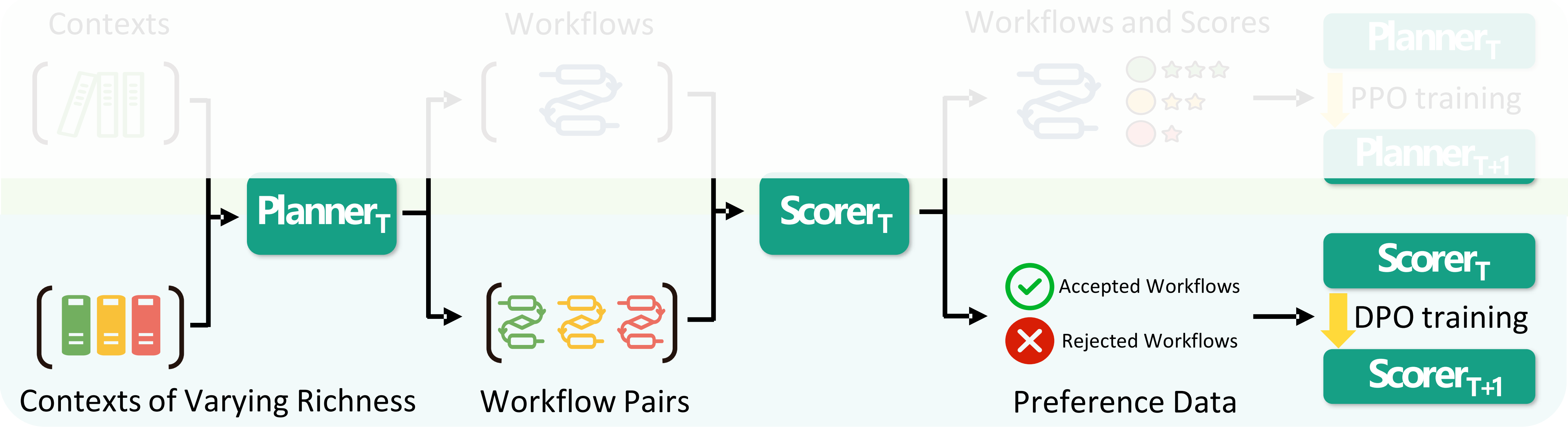
Module #2 Multi-Agent Coevolution

FlowXpert - Module#2



Module #2 PPO for Planner

FlowXpert - Module#2



Module #2 DPO for Scorer

Outline

What are the typical resolutions for a cloud incident?

➔ Workflow, step-by-step guidance and executable scripts

How to transform a naive LLM into a workflow generator in the field of cloud services?

➔ Support of domain knowledge, alignment of application capability

Framework design

➔ Knowledge Base Construction, Multi-Agent Coevolution

Evaluation

➔ Benchmark tests, online deployment, case study

Evaluation: Dataset and Metric

Evaluation Dataset:

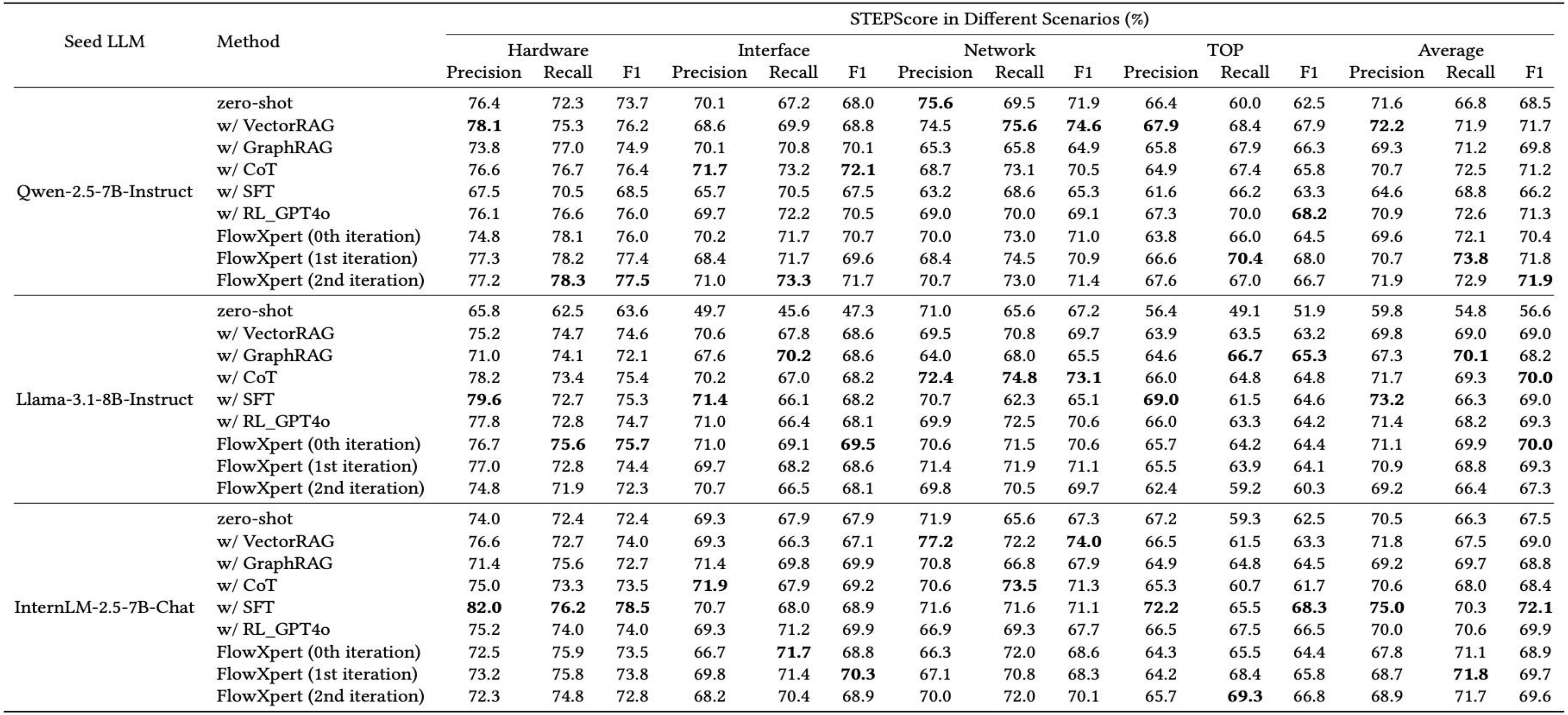
- From operation documents of Huawei Cloud's datacenter network (DCN team)
- 252 user queries and their corresponding standard workflows
- 4 domains: hardware, interface, network, top

Metric:

- We propose STEPScore as a tailored metric
- The *Precision* indicates how closely the generated steps match the standard steps.
- The *Recall* indicates how well the standard steps are retrieved in the generated steps.

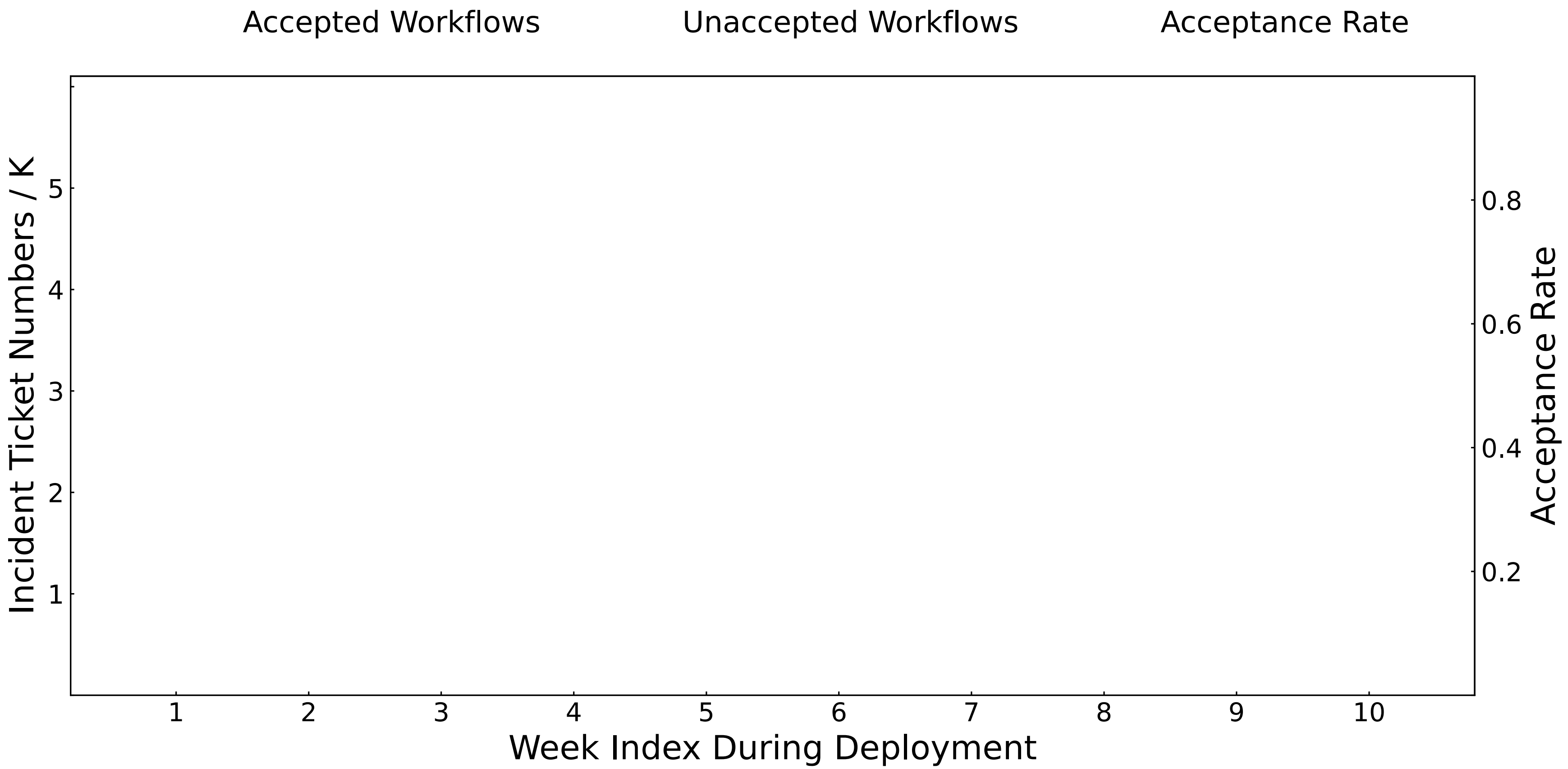
$$Precision = \frac{1}{|S_g|}\sum_{s_i \in S_g} max_{s_j \in S_r} \cos(E(S_i), E(S_j)) \qquad Recall = \frac{1}{|S_r|}\sum_{s_j \in S_r} max_{s_i \in S_g} \cos(E(S_i), E(S_j))$$

Evaluation: Overall Performance



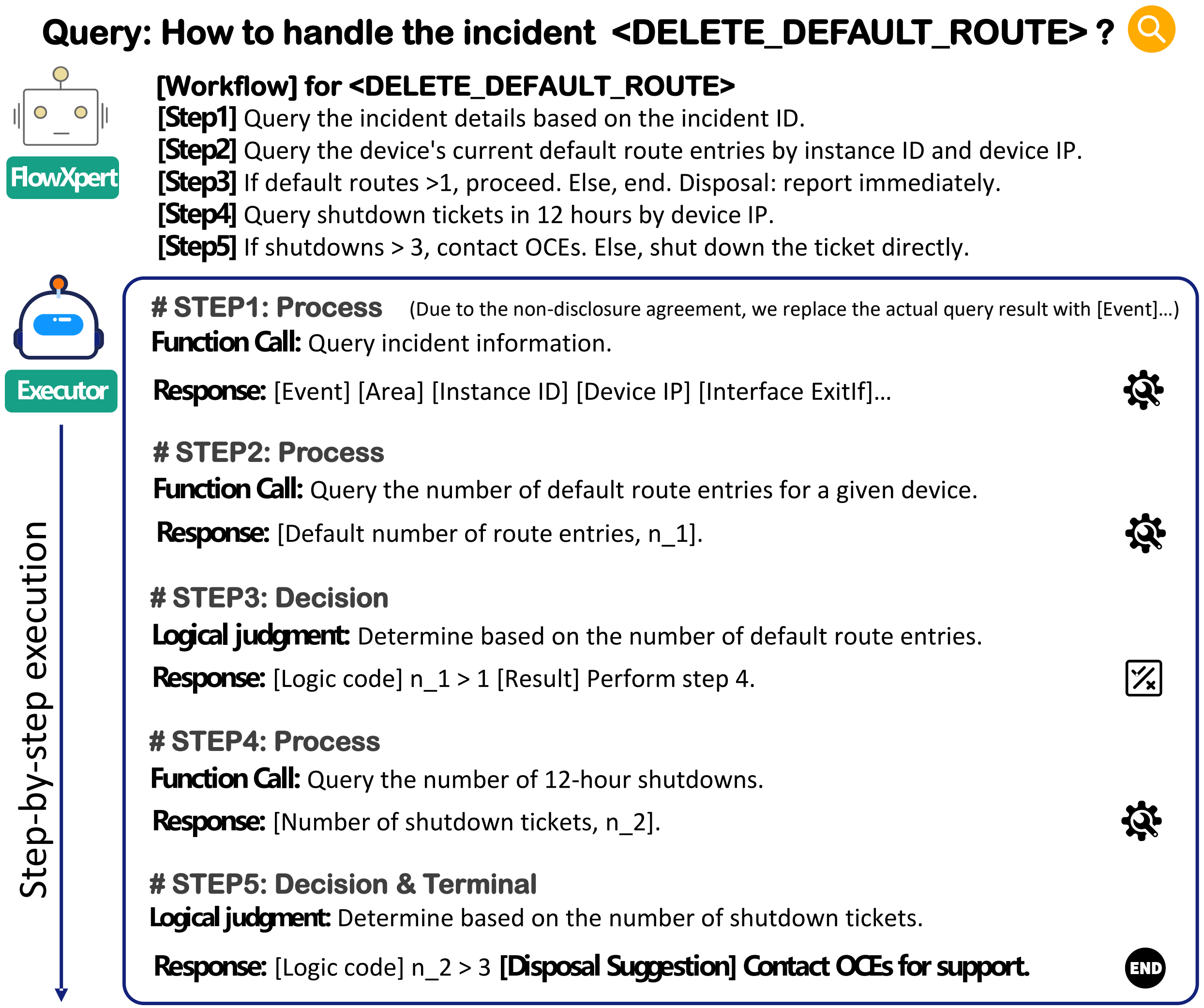
Seed LLM	Method	STEPScore in Different Scenarios (%)														
		Hardware			Interface			Network			TOP			Average		
		Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Qwen-2.5-7B-Instruct	zero-shot	76.4	72.3	73.7	70.1	67.2	68.0	**75.6**	69.5	71.9	66.4	60.0	62.5	71.6	66.8	68.5
	w/ VectorRAG	**78.1**	75.3	76.2	68.6	69.9	68.8	74.5	**75.6**	**74.6**	**67.9**	68.4	67.9	**72.2**	71.9	71.7
	w/ GraphRAG	73.8	77.0	74.9	70.1	70.8	70.1	65.3	65.8	64.9	65.8	67.9	66.3	69.3	71.2	69.8
	w/ CoT	76.6	76.7	76.4	**71.7**	73.2	**72.1**	68.7	73.1	70.5	64.9	67.4	65.8	70.7	72.5	71.2
	w/ SFT	67.5	70.5	68.5	65.7	70.5	67.5	63.2	68.6	65.3	61.6	66.2	63.3	64.6	68.8	66.2
	w/ RL_GPT4o	76.1	76.6	76.0	69.7	72.2	70.5	69.0	70.0	69.1	67.3	70.0	**68.2**	70.9	72.6	71.3
	FlowXpert (0th iteration)	74.8	78.1	76.0	70.2	71.7	70.7	70.0	73.0	71.0	63.8	66.0	64.5	69.6	72.1	70.4
	FlowXpert (1st iteration)	77.3	78.2	77.4	68.4	71.7	69.6	68.4	74.5	70.9	66.6	**70.4**	68.0	70.7	**73.8**	71.8
	FlowXpert (2nd iteration)	77.2	**78.3**	**77.5**	71.0	**73.3**	71.7	70.7	73.0	71.4	67.6	67.0	66.7	71.9	72.9	**71.9**
Llama-3.1-8B-Instruct	zero-shot	65.8	62.5	63.6	49.7	45.6	47.3	71.0	65.6	67.2	56.4	49.1	51.9	59.8	54.8	56.6
	w/ VectorRAG	75.2	74.7	74.6	70.6	67.8	68.6	69.5	70.8	69.7	63.9	63.5	63.2	69.8	69.0	69.0
	w/ GraphRAG	71.0	74.1	72.1	67.6	**70.2**	68.6	64.0	68.0	65.5	64.6	**66.7**	**65.3**	67.3	**70.1**	68.2
	w/ CoT	78.2	73.4	75.4	70.2	67.0	68.2	**72.4**	**74.8**	**73.1**	66.0	64.8	64.8	71.7	69.3	**70.0**
	w/ SFT	**79.6**	72.7	75.3	**71.4**	66.1	68.2	70.7	62.3	65.1	**69.0**	61.5	64.6	**73.2**	66.3	69.0
	w/ RL_GPT4o	77.8	72.8	74.7	71.0	66.4	68.1	69.9	72.5	70.6	66.0	63.3	64.2	71.4	68.2	69.3
	FlowXpert (0th iteration)	76.7	**75.6**	**75.7**	71.0	69.1	**69.5**	70.6	71.5	70.6	65.7	64.2	64.4	71.1	69.9	**70.0**
	FlowXpert (1st iteration)	77.0	72.8	74.4	69.7	68.2	68.6	71.4	71.9	71.1	65.5	63.9	64.1	70.9	68.8	69.3
	FlowXpert (2nd iteration)	74.8	71.9	72.3	70.7	66.5	68.1	69.8	70.5	69.7	62.4	59.2	60.3	69.2	66.4	67.3
InternLM-2.5-7B-Chat	zero-shot	74.0	72.4	72.4	69.3	67.9	67.9	71.9	65.6	67.3	67.2	59.3	62.5	70.5	66.3	67.5
	w/ VectorRAG	76.6	72.7	74.0	69.3	66.3	67.1	**77.2**	72.2	**74.0**	66.5	61.5	63.3	71.8	67.5	69.0
	w/ GraphRAG	71.4	75.6	72.7	71.4	69.8	69.9	70.8	66.8	67.9	64.9	64.8	64.5	69.2	69.7	68.8
	w/ CoT	75.0	73.3	73.5	**71.9**	67.9	69.2	70.6	**73.5**	71.3	65.3	60.7	61.7	70.6	68.0	68.4
	w/ SFT	**82.0**	**76.2**	**78.5**	70.7	68.0	68.9	71.6	71.6	71.1	**72.2**	65.5	**68.3**	**75.0**	70.3	**72.1**
	w/ RL_GPT4o	75.2	74.0	74.0	69.3	71.2	69.9	66.9	69.3	67.7	66.5	67.5	66.5	70.0	70.6	69.9
	FlowXpert (0th iteration)	72.5	75.9	73.5	66.7	**71.7**	68.8	66.3	72.0	68.6	64.3	65.5	64.4	67.8	71.1	68.9
	FlowXpert (1st iteration)	73.2	75.8	73.8	69.8	71.4	**70.3**	67.1	70.8	68.3	64.2	68.4	65.8	68.7	**71.8**	69.7
	FlowXpert (2nd iteration)	72.3	74.8	72.8	68.2	70.4	68.9	70.0	72.0	70.1	65.7	**69.3**	66.8	68.9	71.7	69.6

Online Deployment: For OCEs



For OCEs: (1) A 10-week deployment in Huawei Cloud DCN, 189 common incident types, 34K incident tickets; (2) average 22.1 seconds for a workflow v.s. 7-hour effort of a 7-person OCE team, including 2 experts.

Case Study: For AI Executors



Query: How to handle the incident <DELETE_DEFAULT_ROUTE> ?

[Workflow] for <DELETE_DEFAULT_ROUTE>
[Step1] Query the incident details based on the incident ID.
[Step2] Query the device's current default route entries by instance ID and device IP.
[Step3] If default routes >1, proceed. Else, end. Disposal: report immediately.
[Step4] Query shutdown tickets in 12 hours by device IP.
[Step5] If shutdowns > 3, contact OCEs. Else, shut down the ticket directly.

# STEP1: Process (Due to the non-disclosure agreement, we replace the actual query result with [Event]...)
Function Call: Query incident information.
Response: [Event] [Area] [Instance ID] [Device IP] [Interface ExitIf]...

# STEP2: Process
Function Call: Query the number of default route entries for a given device.
Response: [Default number of route entries, n_1].

# STEP3: Decision
Logical judgment: Determine based on the number of default route entries.
Response: [Logic code] n_1 > 1 [Result] Perform step 4.

# STEP4: Process
Function Call: Query the number of 12-hour shutdowns.
Response: [Number of shutdown tickets, n_2].

# STEP5: Decision & Terminal
Logical judgment: Determine based on the number of shutdown tickets.
Response: [Logic code] n_2 > 3 **[Disposal Suggestion] Contact OCEs for support.**

For AI Executors: A case study of autonomous AI Executor for incident handling.

Future Directions

- **Novel Incident Handling:** …
- **Quality of Retrieved Knowledge:** …
- **Execution Constraints:** …

Takeaways

- ✓ We propose FlowXpert, a framework that orchestrates troubleshooting workflows by **integrating domain knowledge support and aligned knowledge application.** Specifically, we (1) define a domain ontology to guide the knowledge base construction, (2) implement multi-agent coevolution through PPO and DPO tuning, (3) design a preference data synthesis method controlled by contextual richness.
- ✓ We introduce **STEPScore**, a metric designed around core characteristics of workflows, and conduct extensive benchmark tests based on real-world incidents from Huawei Cloud DCN team.
- ✓ In production of DCN, our framework **contributed a lot to both**

Thank you!

- Reporter: Binpeng Shi
- Email: shibinpeng23@mail.nankai.edu.cn