ICLR
International Conference On Learning Representations

浙江大学 ZHEJIANG UNIVERSITY

Alibaba

Benchmarking Agentic Workflow Generation

Shuofei Qiao♠*, Runnan Fang♠*, Zhisong Qiu♠*, Xiaobin Wang♢,

Ningyu Zhang♠†, Yong Jiang♢†, Pengjun Xie♢, Fei Huang♢, Huajun Chen♠♡†

♠Zhejiang University ♢Alibaba Group

♡Zhejiang Key Laboratory of Big Data Intelligent Computing

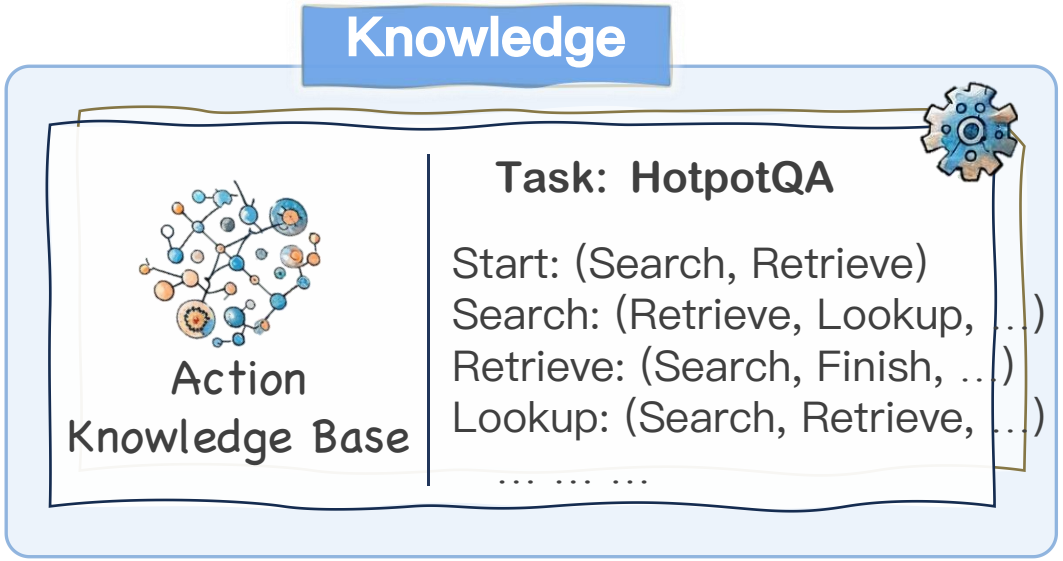
shuofei@zju.edu.cn

Shuofei Qiao

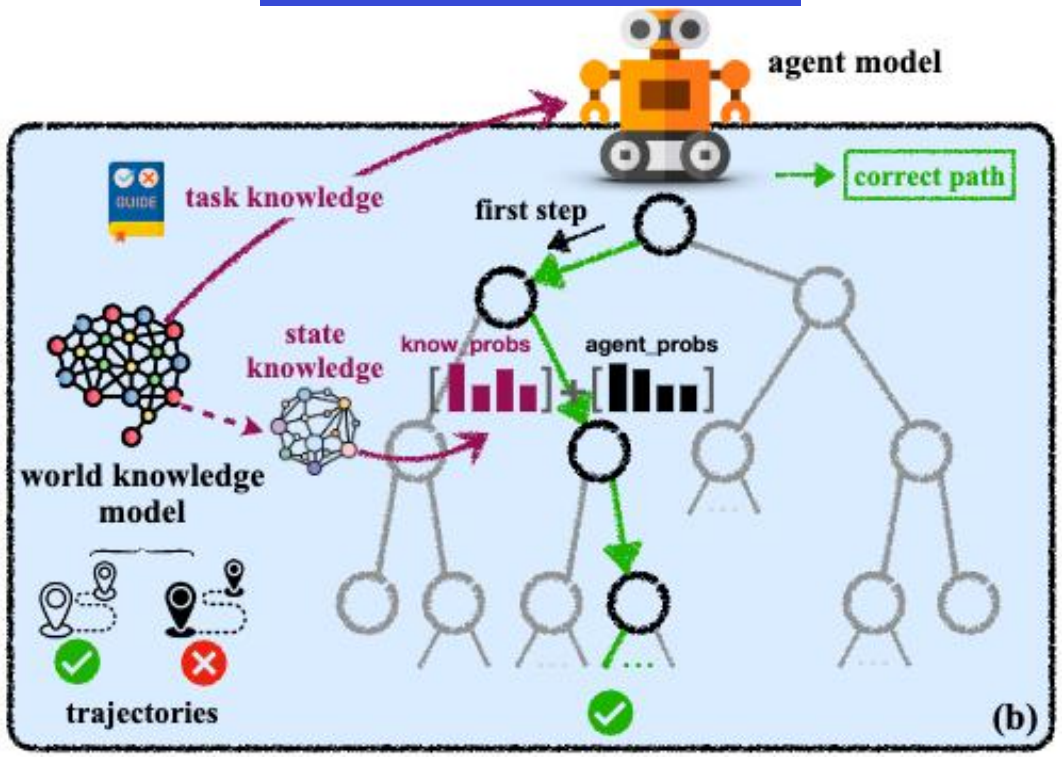
03/12/2025

Motivation

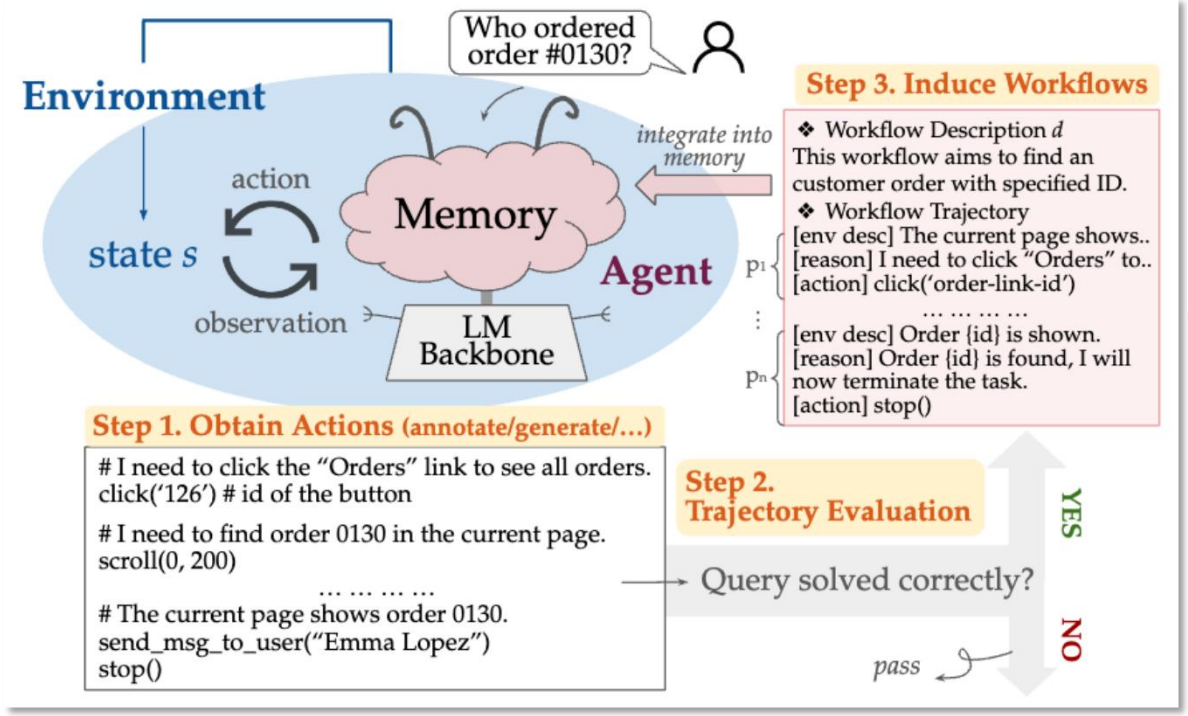
Symbolic Knowledge

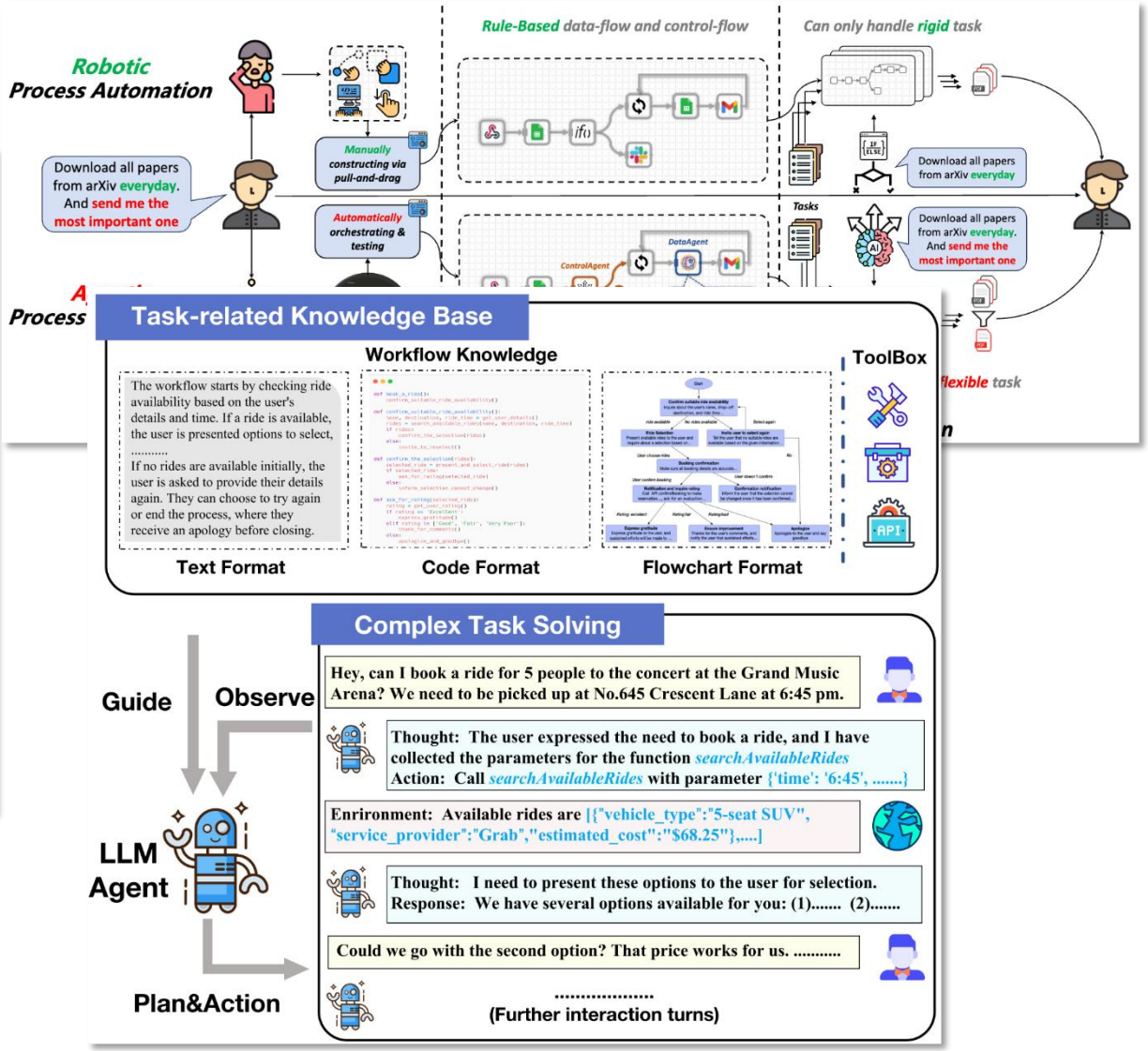


Parameterized Knowledge



Motivation



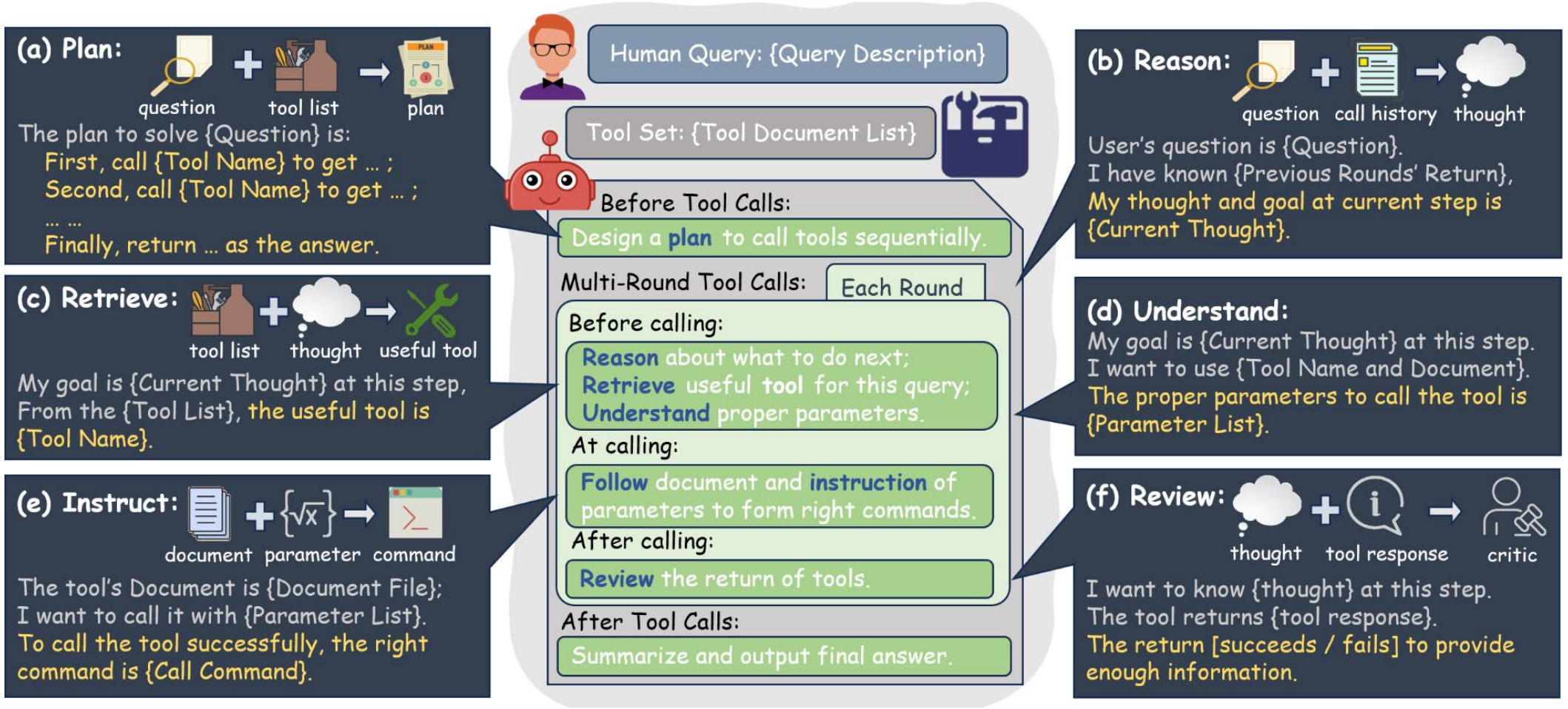


[1] Agent Workflow Memory (2024)
[2] ProAgent: From Robotic Process Automation to Agentic Process Automation (2023)
[3] FlowBench: Revisiting and Benchmarking Workflow-Guided Planning for LLM-based Agents (2024)

Motivation

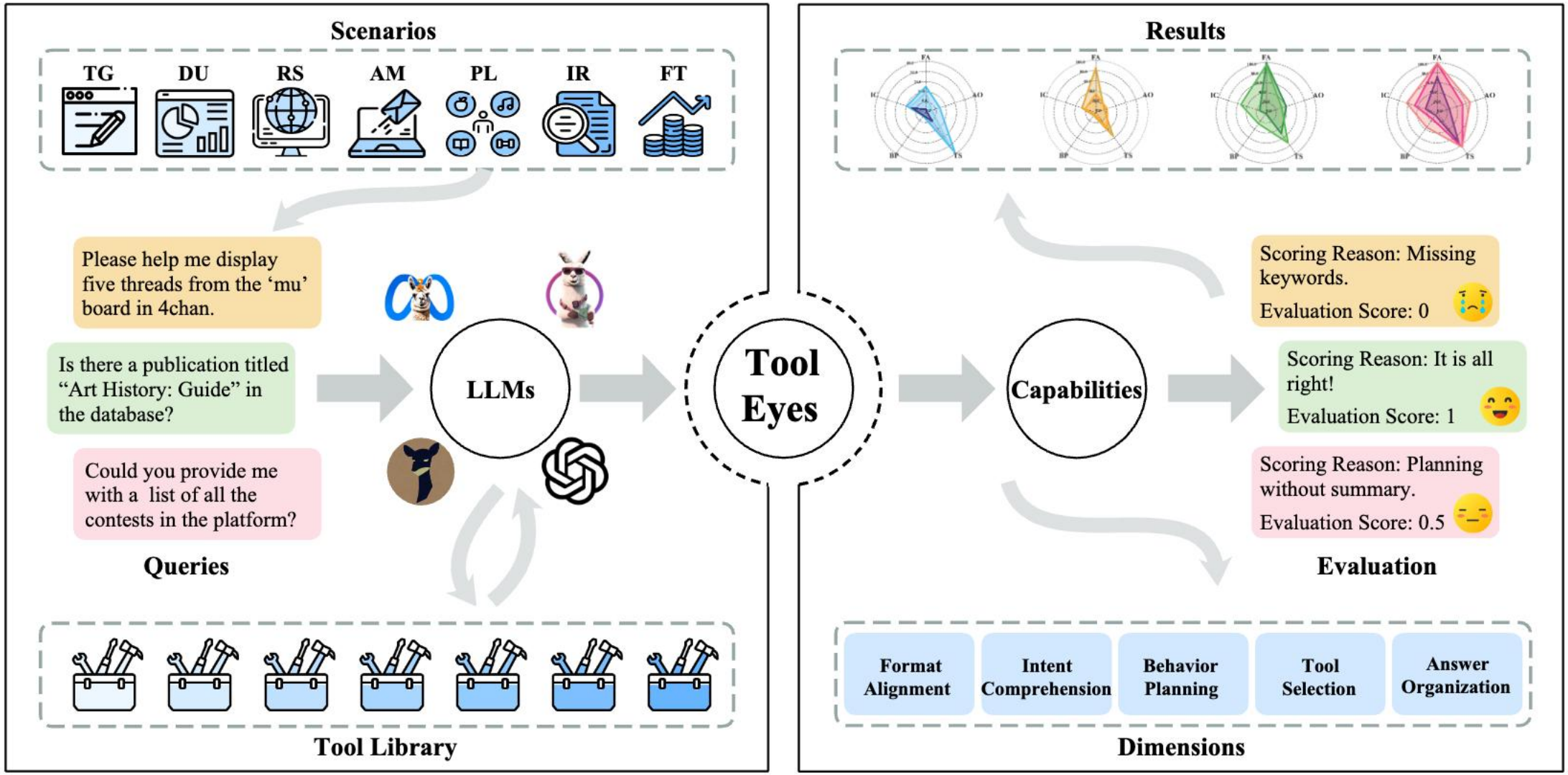
Limited scope of scenarios.

Sole emphasis on linear relationships between subtasks.

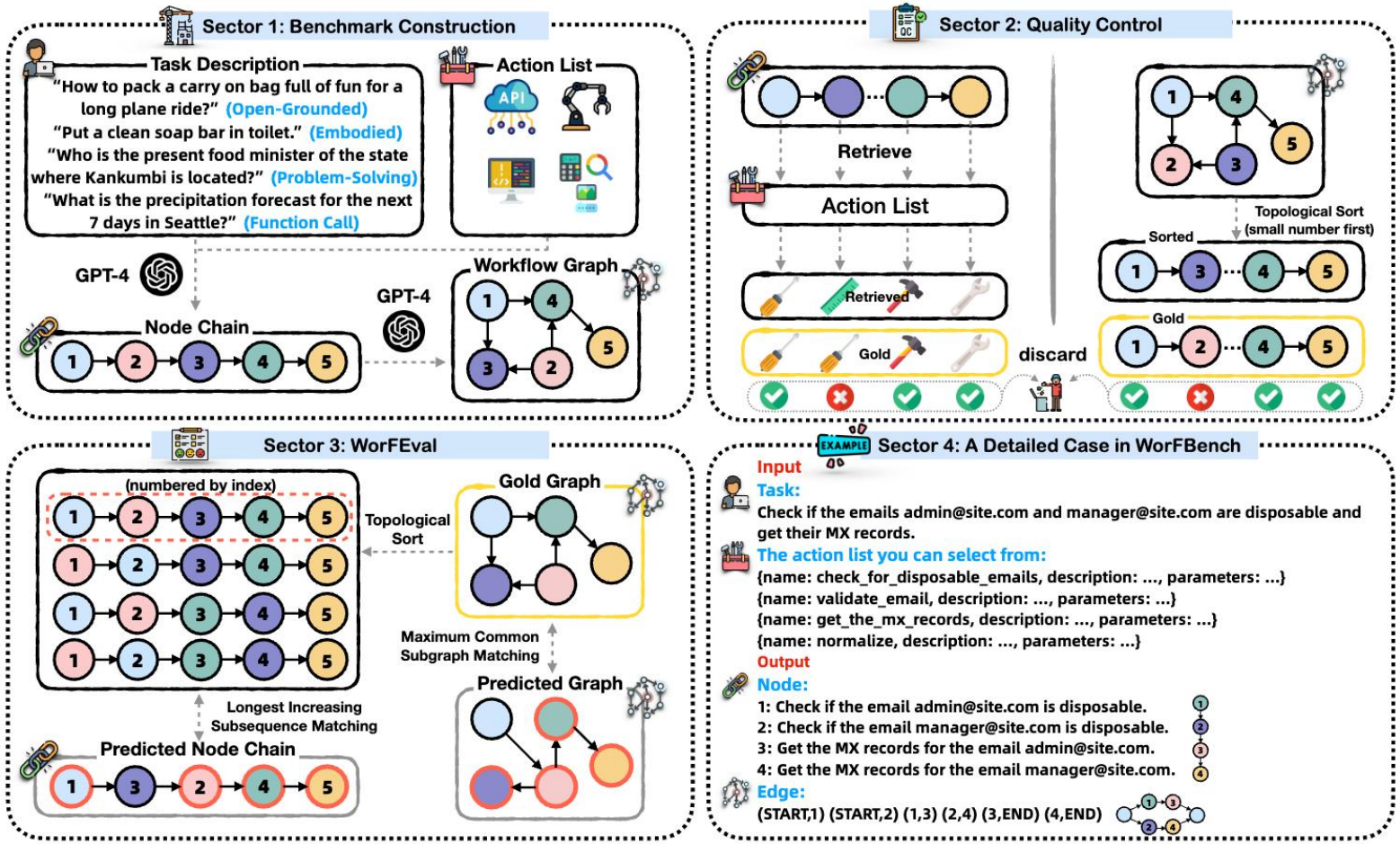


Motivation

Evaluations heavily rely on GPT-3.5/4

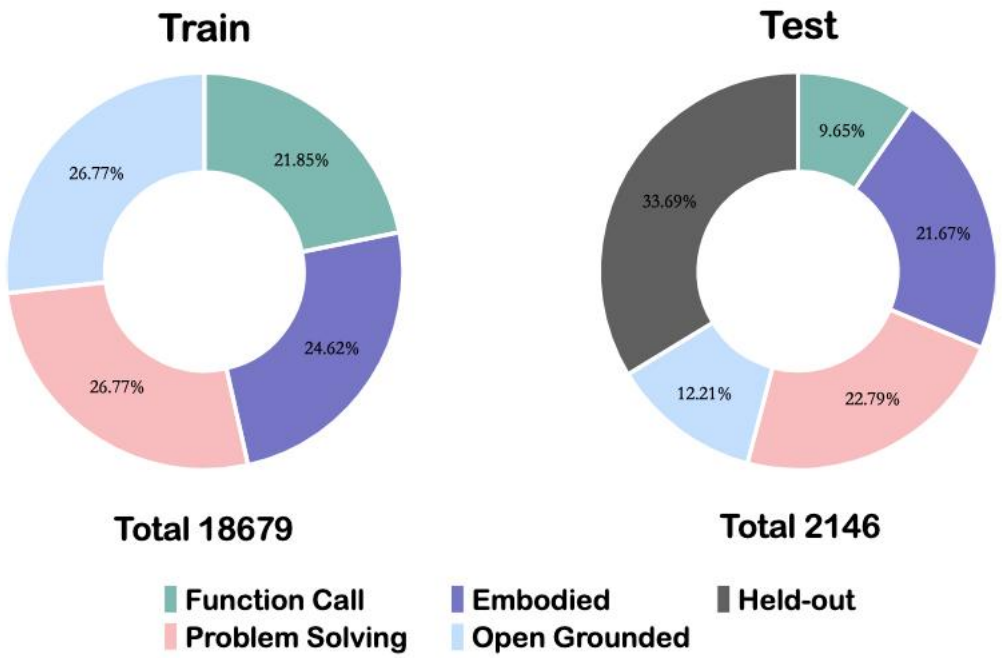


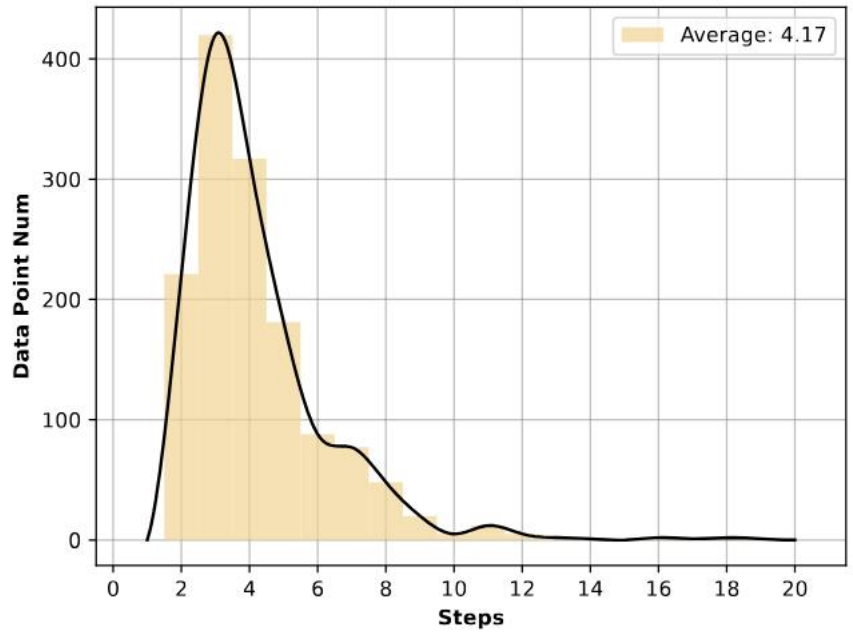
WorfBench



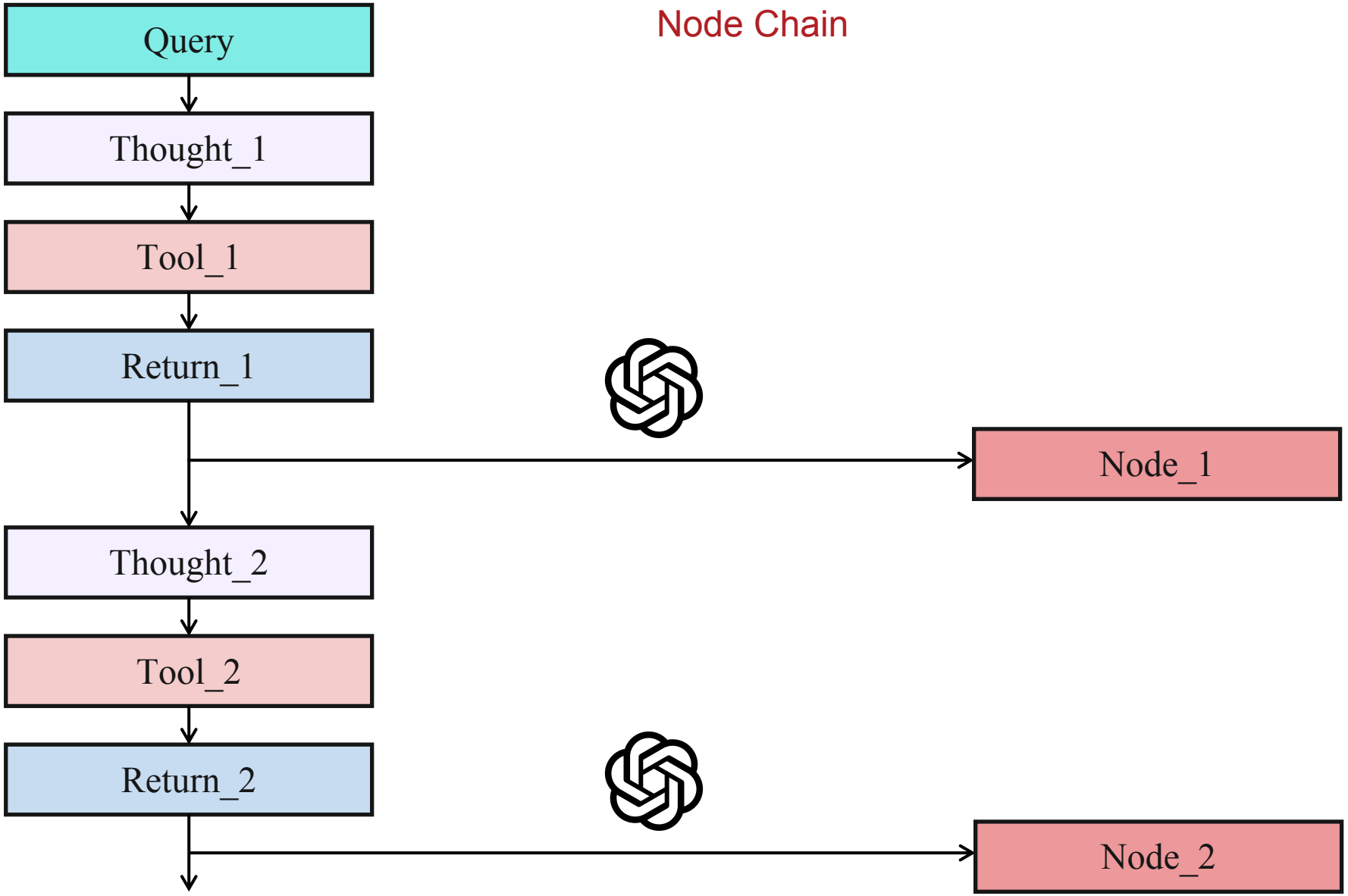
- Multi-faceted scenarios.
- Complex workflow structures.
- Strict quality control and data filtering.
- Accurate quantitative evaluation (WorfEval).

WorfBench


Train
21.85%
24.62%
26.77%
26.77%
Total 18679
Test
9.65%
21.67%
22.79%
12.21%
33.69%
Total 2146
Function Call
Embodied
Held-out
Problem Solving
Open Grounded


Average: 4.17
Data Point Num
Steps
0
100
200
300
400
0
2
4
6
8
10
12
14
16
18
20

Dataset Construction



Dataset Construction

"query": "I am conducting a research project on basketball statistics and I require the player season stats for Jayson Tatum, the top 20 players with the most assists in the 2023 season, and the top scorers in the 2011 playoffs. Can you assist me by providing this information?",

"node_step": 1,

"node_task": "Search player season stats for Jayson Tatum."

"node_step": 2,

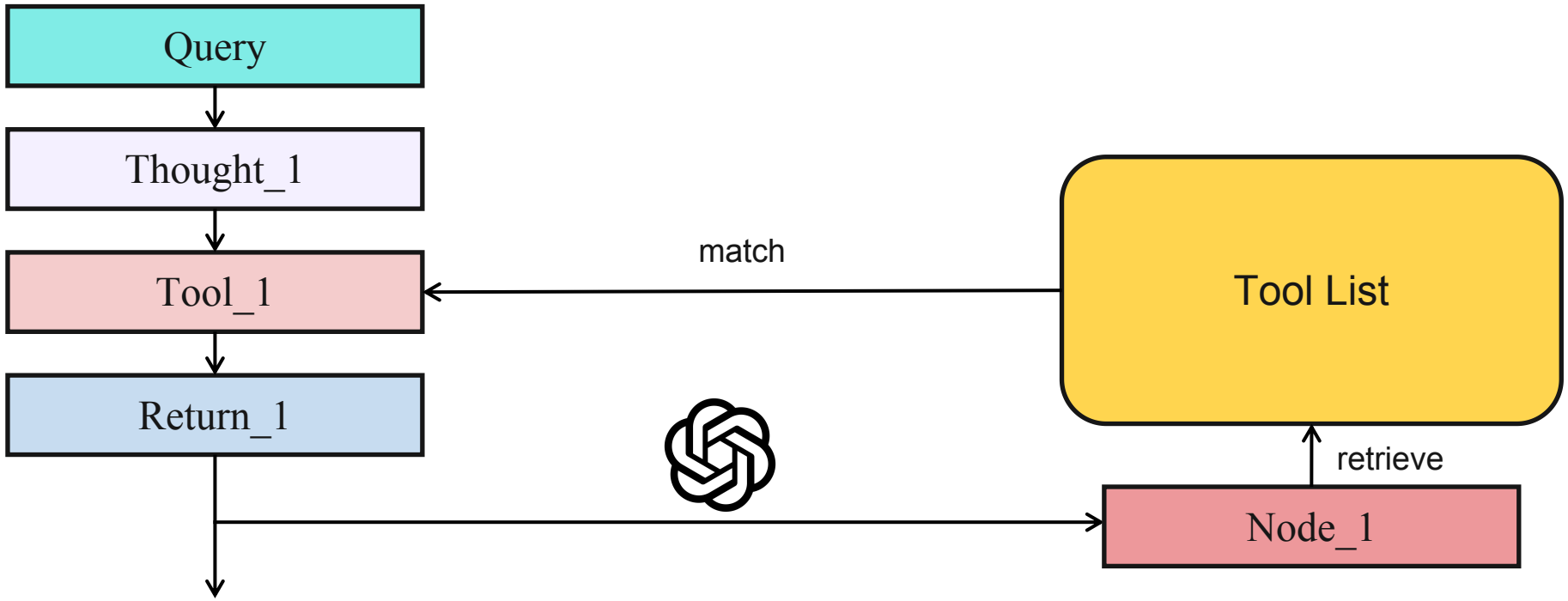
"node_task": "Find the top 20 players with the most assists in the 2023 season."

"node_step": 3,

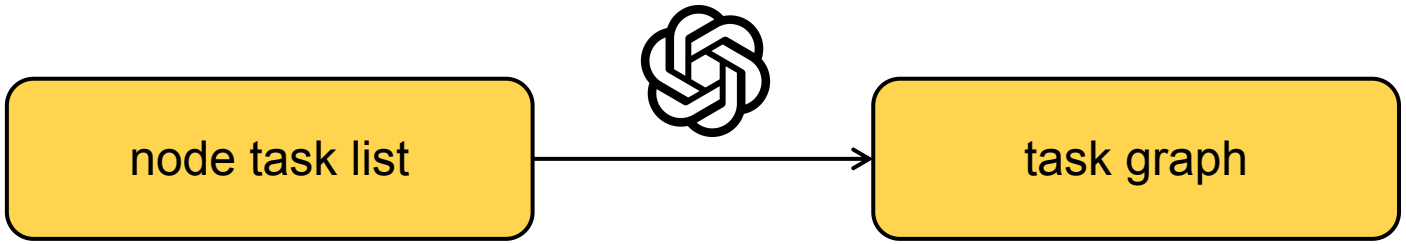
"node_task": "Retrieve the top scorers in the 2011 playoffs."

Dataset Construction

node chain list quality control



Dataset Construction



Node:

1: Combine the brine ingredients.

2: Brine the chicken wings for 1 hour.

3: Rinse and dry the wings.

4: Preheat the grill for indirect heat.

5: Toss the wings in oil.

6: Sprinkle with lemon pepper seasoning.

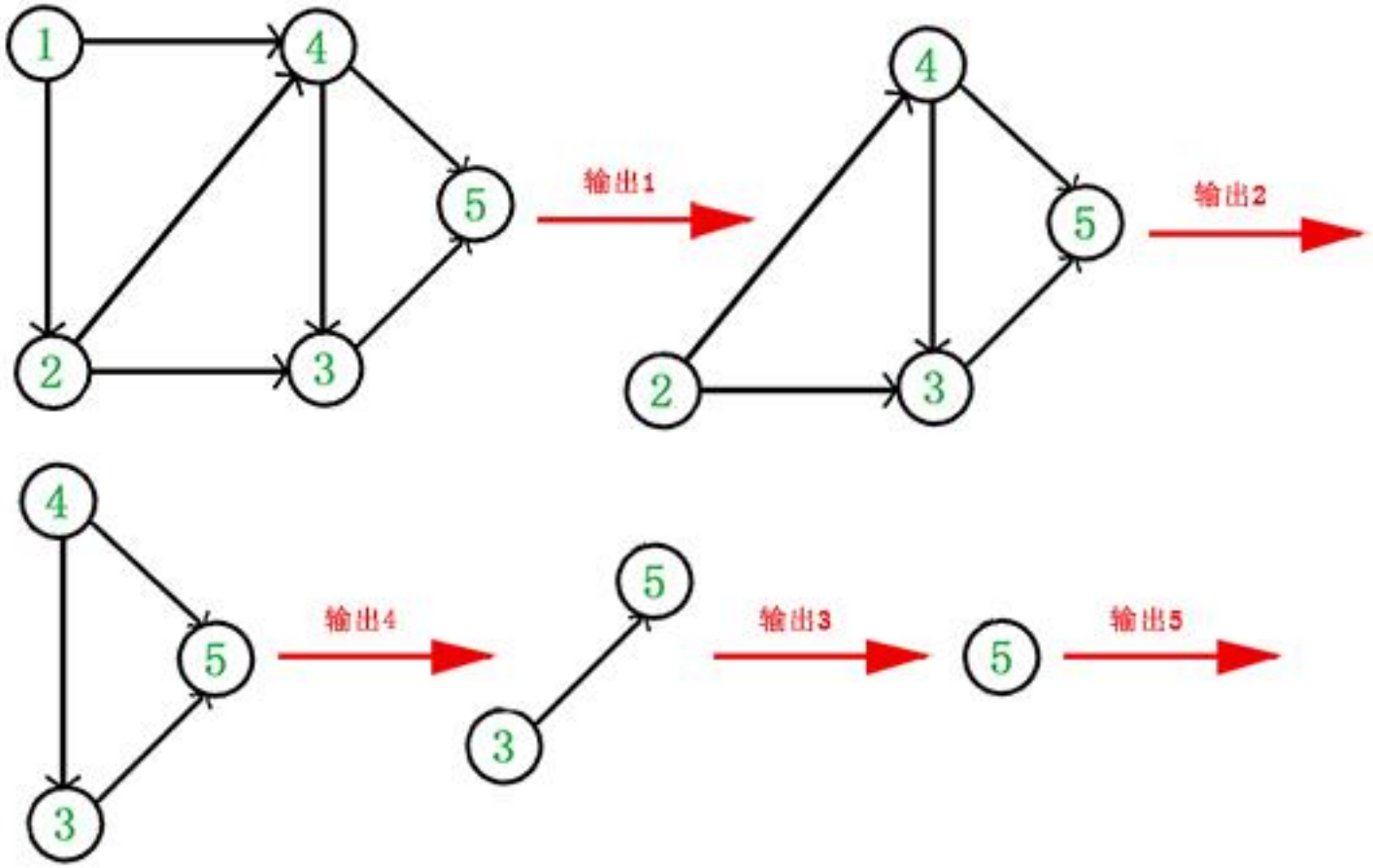
7: Cook for 45 to 60 minutes.

8: Serve immediately.

Edge: (START,1) (1,2) (2,3) (3,4) (3,5) (4,7) (5,6) (6,7) (7,8) (8,END)

Dataset Construction

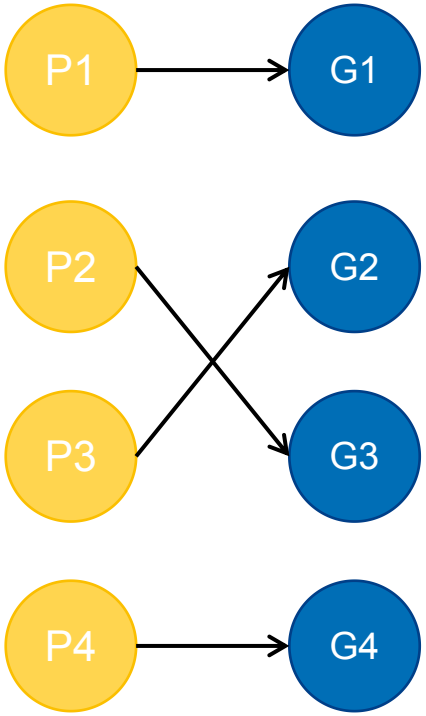
graph quality control



拓扑排序

WorfEval

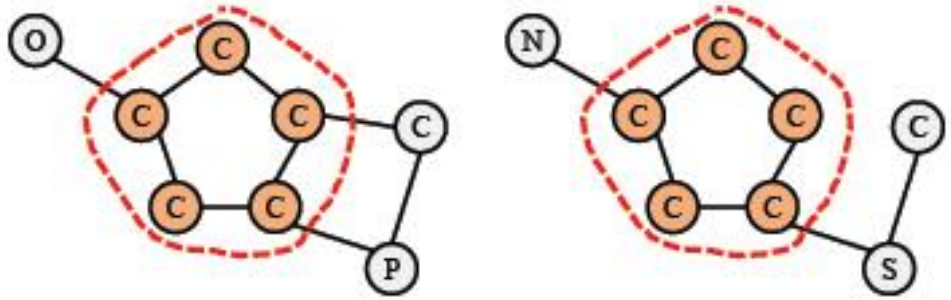
list score



Max-weighted Bipartite Matching Algorithm

Longest Increasing Subsequence

graph score



Maximum Common Induced Subgraph

Experiments

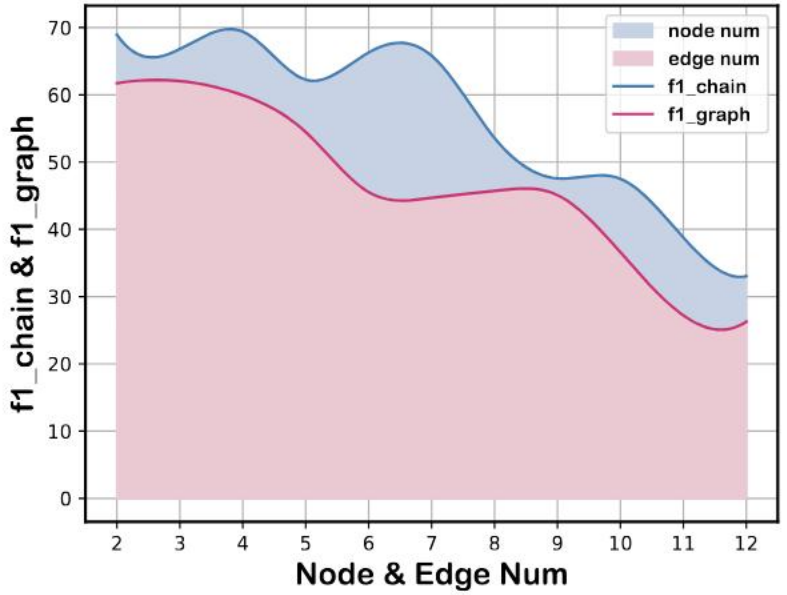
Model	Function Call		Problem-Solving		Embodied		Open-Grounded		Average	
	$f1_{\text{chain}}$	$f1_{\text{graph}}$	$f1_{\text{chain}}$	$f1_{\text{graph}}$	$f1_{\text{chain}}$	$f1_{\text{graph}}$	$f1_{\text{chain}}$	$f1_{\text{graph}}$	$f1_{\text{chain}}$	$f1_{\text{graph}}$
Closed-Sourced										
Claude-3.5	66.44	55.06	67.28	55.50	**71.74**	**56.71**	**61.33**	**42.88**	66.70	**52.53**
GPT-3.5	73.36	60.32	69.86	54.50	64.57	49.17	47.67	28.10	63.86	48.02
GPT-4	**74.87**	**62.11**	67.18	55.24	70.94	56.17	56.30	36.36	**67.32**	52.47
O1-preview	70.68	57.11	**72.76**	**59.25**	69.90	54.19	53.47	35.97	66.70	51.63
Open-Sourced										
GLM-4-9B	59.27	36.34	58.91	40.15	53.17	36.15	44.04	22.56	53.85	33.80
Phi-3-small	57.66	40.71	55.76	39.75	54.77	37.52	**44.65**	22.66	53.21	35.16
Llama-3.1-8B	63.30	43.62	64.49	46.79	56.23	36.40	44.58	**25.48**	57.15	38.08
Mistral-7B	67.30	51.67	61.27	45.35	64.59	**48.83**	40.97	21.48	58.53	41.83
Qwen-2-7B	**70.79**	**55.50**	68.65	52.13	62.83	46.25	39.29	20.89	60.39	43.69
InternLM-2.5-7B	68.43	52.99	**72.92**	**57.80**	**65.77**	48.09	40.84	21.27	**61.99**	**45.03**
Llama-2-13B	53.32	34.33	53.74	38.69	44.27	30.55	37.17	23.14	47.12	31.68
WizardLM-13B	55.78	36.94	65.42	49.71	55.41	37.34	37.23	21.66	53.46	36.41
Vicuna-13B	53.75	37.66	64.58	50.25	57.99	42.61	38.93	23.11	53.81	38.41
Qwen-1.5-14B	65.73	46.86	58.80	43.89	60.55	44.14	41.73	21.44	56.70	39.08
Phi-3-medium	**67.71**	**47.26**	**71.15**	**54.85**	**65.11**	**49.99**	**42.73**	**23.77**	**61.68**	**43.97**
WizardLM-70B	63.47	45.46	63.92	47.93	59.15	42.87	45.27	26.89	57.95	40.79
Mixtral-8×7B	66.13	48.83	**71.89**	57.58	72.08	54.94	42.96	23.21	63.26	46.14
Llama-3.1-70B	64.41	**52.72**	70.37	57.05	69.98	55.52	**53.64**	**33.06**	64.60	49.59
Qwen-2-72B	**71.67**	52.31	70.63	**58.13**	**73.24**	**58.49**	53.43	32.89	**67.24**	**50.46**

① Which is more challenging for LLM agents, linear planning or graph planning?

② How is Scaling Law manifested in workflow generation?

Experiments

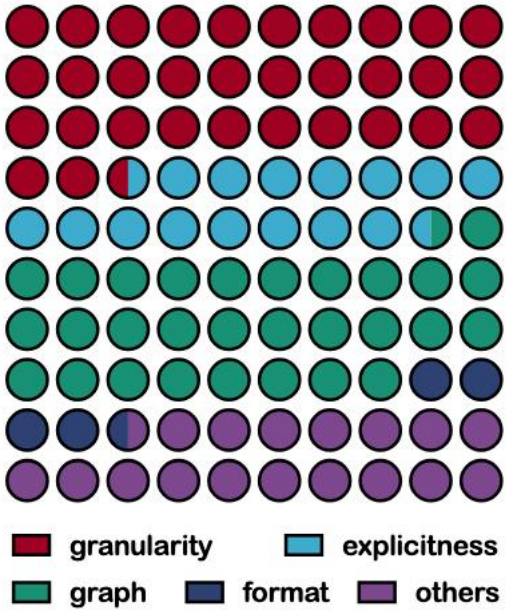
③ How far are existing LLM agents from being real workflow planning experts?



Model	Held-in Tasks		Held-out Tasks			
	Average		Seal-Tools		InterCodeSQL	
	f_{chain}	f_{graph}	f_{chain}	f_{graph}	f_{chain}	f_{graph}
GPT-3.5	63.86	48.02	95.91	76.63	<u>65.30</u>	53.07
GPT-4	67.32	52.47	**96.58**	80.25	**66.35**	**54.36**
Qwen-2-7B	60.39	43.69	92.68	74.75	54.20	39.72
InternLM-2.5-7B	61.99	45.03	93.07	74.43	55.06	42.20
Phi-3-medium	61.68	43.97	94.11	79.45	58.45	46.62
Llama-3.1-70B	64.60	49.59	94.40	80.11	63.49	<u>53.66</u>
Qwen-2-72B	67.24	50.46	94.47	78.90	63.86	52.47
Qwen-2-7B+FT	**79.35**	**70.38**	<u>96.49</u>	<u>82.82</u>	62.37	48.72
InternLM-2.5-7B+FT	<u>78.98</u>	<u>69.33</u>	95.83	**83.72**	63.78	50.97

Experiments

④ What shall we do to enhance the workflow generation capability of LLM agents?



- **Granularity** pertains to a deficiency in prior knowledge of the environment.
- **Explicitness** reflects a lack of understanding of the environmental task, resulting in insufficiently specific subtasks.
- **Graph** issues arise from a lack of comprehension regarding the dependencies of environmental actions

World Knowledge & World Model

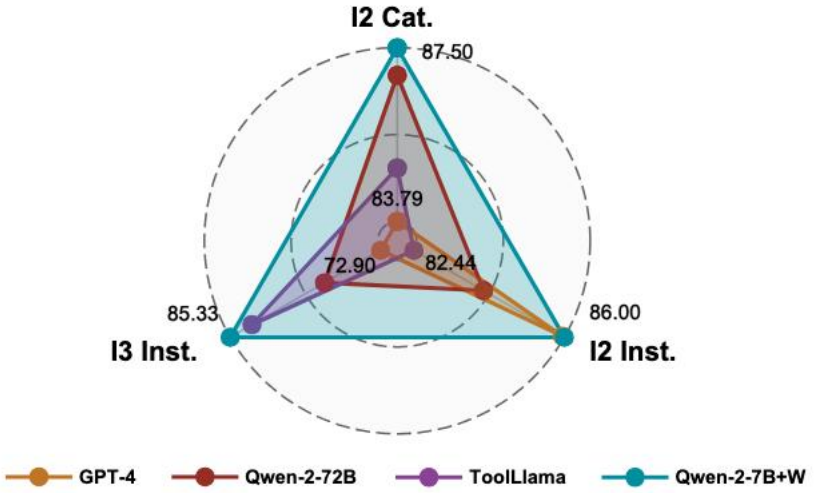
The Role of Workflow for Agent Planning

Enhance End-to-end Performance

- Workflow as Structured Prior Knowledge.

Model	ALFWorld		WebShop
	seen	unseen	
GPT-4	27.14	28.36	55.62
GPT-4+W	**40.71** ↑13.57	**47.01** ↑18.65	**56.49** ↑0.87
Llama-3.1-8B	1.49	5.00	51.03
Llama-3.1-8B+W	**8.21** ↑6.72	**12.14** ↑7.14	**52.28** ↑1.25
Qwen-2-72B	53.57	56.72	58.95
Qwen-2-72B+W	**56.43** ↑2.86	**62.29** ↑5.57	**60.55** ↑1.60

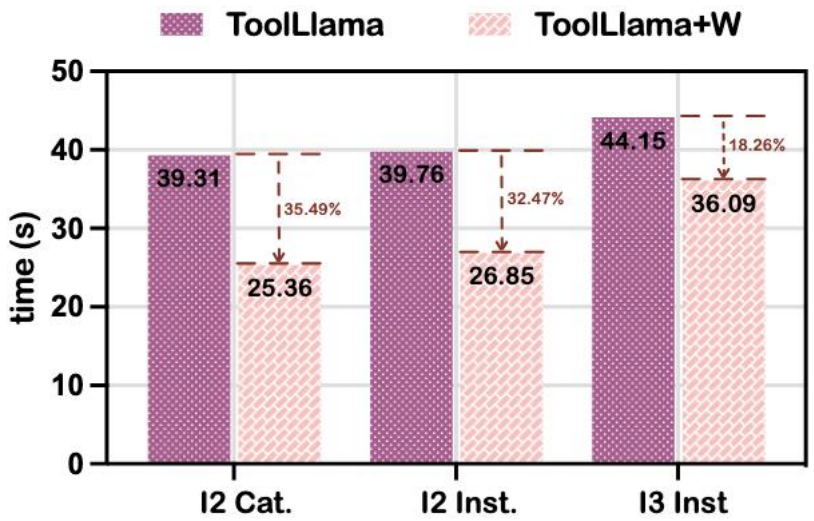
- Workflow as CoT Augmentation.



The Role of Workflow for Agent Planning

Reduce End-to-end Inference-time

- Parallel Planning Steps.



- Shorten Planning Steps.

Model	ALFWorld		WebShop
	seen	unseen	
GPT-4	17.19	17.43	5.80
GPT-4+W	**15.64** ↓1.55	**15.85** ↓1.58	**5.72** ↓0.08
Llama-3.1-8B	19.81	19.43	7.08
Llama-3.1-8B+W	**19.09** ↓0.72	**18.38** ↓1.05	**6.77** ↓0.31
Qwen-2-72B	14.39	14.67	3.88
Qwen-2-72B+W	**14.05** ↓0.34	**13.94** ↓0.73	**3.73** ↓0.15

Conlusion & Limitations

- WorFBench: multi-faceted, complex workflow structure, accurate quantitative evaluation (WorfEval)
- We conduct comprehensive evaluation on various closed-sourced and open-sourced models with different scales. We further exploit the generated workflows to facilitate downstream tasks and achieve superior and efficient performance.

Limitations

- Code workflow (PDDL)
- Iterative generation
- Node uncertainty

浙江大學
ZHEJIANG UNIVERSITY

Thank You !