OS智能体原子任务到复合任务的能力泛化研究与系统调度方法

▲ 郭源 上海交通大学

▲ 2025.5.29

目录

CONTENTS

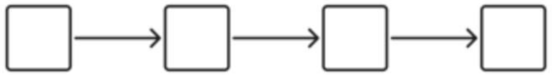
- 研究背景
- UI-NEXUS测试基准
- Agent-NEXUS调度系统
- 实验分析
- 未来展望

01

研究背景

从简单有序到复杂无序任务

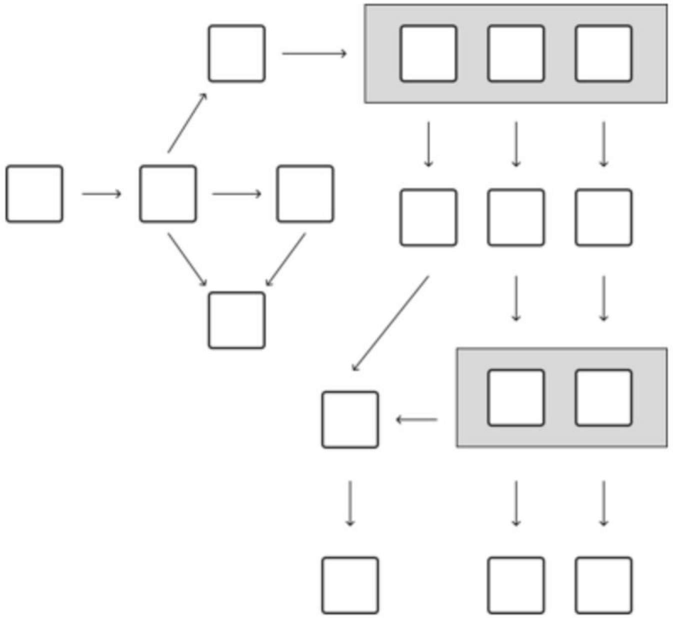
简单有序任务



查下明天上海的天气

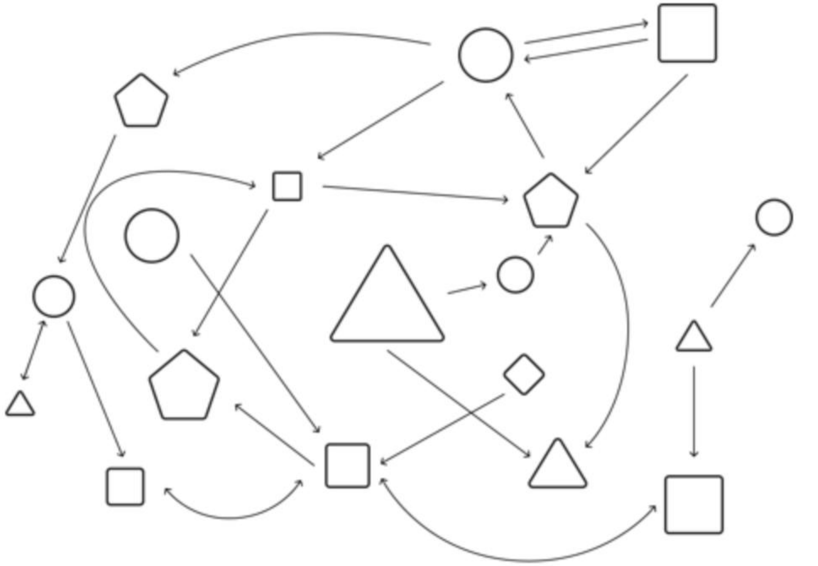
点份昨天晚上的外卖

复杂有序任务



在美团和饿了么分别搜一下肯德基超级
全家桶的价格，并选择更便宜的下单

复杂无序任务



我想申请今年的上海交通大学CS博士项目。
请收集招生信息，在语雀文档写个时间规划
备忘录，并根据我发表的论文方向推荐导师

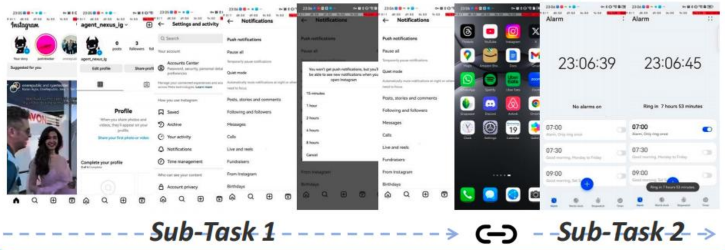
从简单有序到复杂无序任务

真实场景需求驱动的**系统级GUI智能体**，从执行规则明确的简单任务到能胜任复杂有序与复杂无序任务

基于子任务依赖关系的复合指令分类：**拼接型、传递型、深度分析型**

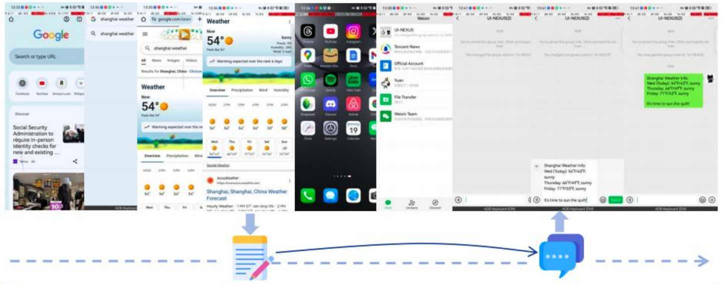
Simple Concatenation

[Instruction] Pause all Instagram notifications for 8 hours, and turn on the clock at 7:00.



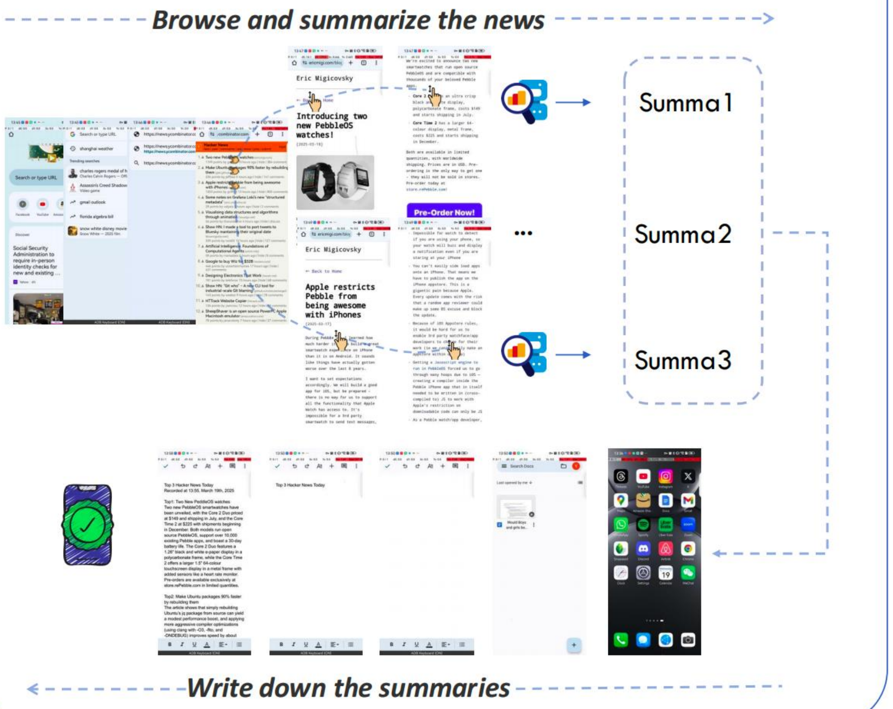
Context Transition

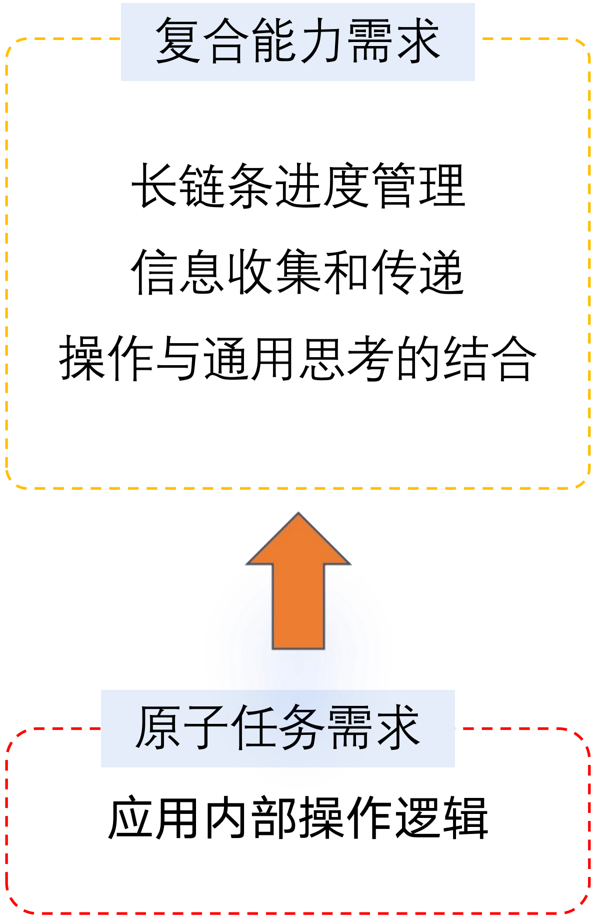
[Instruction] Check Shanghai weather these three days in Chrome. **Send the weather and temperature information** in "UI-NEXUS" WeChat group. If there will be a rainy day, ... And if all the three days are sunny, say ...



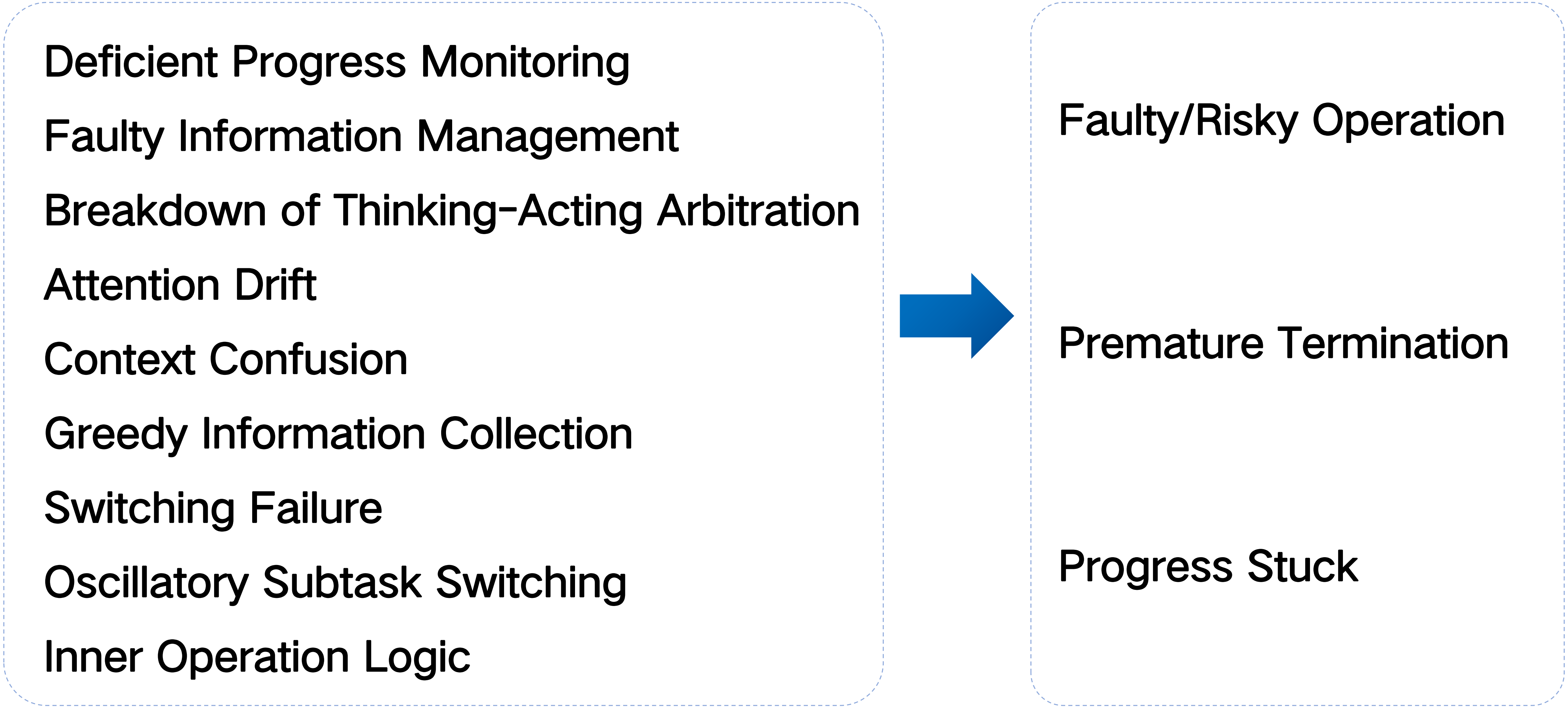
Deep Dive

[Instruction] Use Chrome to visit https://news.ycombinator.com/. **Read** the top three news articles and **summarize each in no more than five concise sentences**. Then, create a file named 'Top 3 Hacker News Today' in Google Docs and **list these three summaries**.


Browse and summarize the news
Introducing two new PebbleOS watches!
Apple restricts Pebble from being awesome with iPhones
Pre-Order Now!
...
Summa1
Summa2
Summa3
Write down the summaries



复合任务的独特挑战



Deficient Progress Monitoring

Faulty Information Management

Breakdown of Thinking-Acting Arbitration

Attention Drift

Context Confusion

Greedy Information Collection

Switching Failure

Oscillatory Subtask Switching

Inner Operation Logic

Faulty/Risky Operation

Premature Termination

Progress Stuck

复合任务的独特挑战

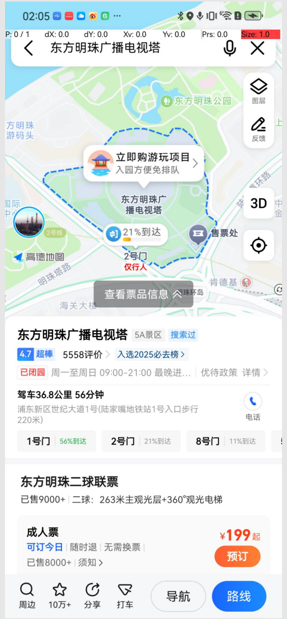
失败案例：注意力涣散，忽略了部分指令要求或整个子任务

User

[Instruction]: Open Gaode Map, search for the Oriental Pearl Radio and Television Tower, then save this address and start the navigation to it. After the navigation starts, go to Settings and set the sound mode to ring in "Sound & Vibration".

Mobile-Agent-V2

Open Gaode and search



东方明珠广播电视塔

东方明珠广播电视塔 5A景区 搜索过

4.7 超棒 5558评价 > 入选2025必去榜 >

已闭园 周一至周日 09:00-21:00 最晚进... | 优待政策 详情 >

驾车36.8公里 56分钟

浦东新区世纪大道1号(陆家嘴地铁站1号入口步行220米)

电话

1号门 56%到达 | 2号门 21%到达 | 8号门 11%到达

东方明珠二球联票

已售9000+ | 二球：263米主观光层+360°观光电梯

成人票 ¥199起

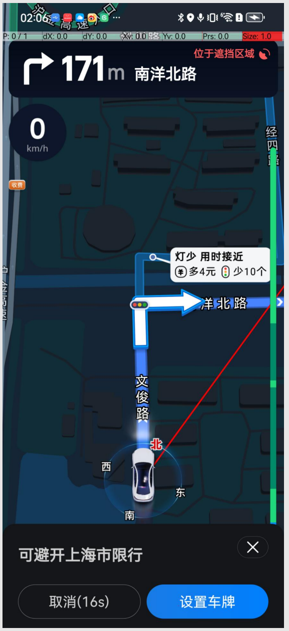
可订今日 | 随时退 | 无需换票 |

已售8000+ | 须知 >

预订

Search for Location ✓

Save the Address ×



171m 南洋北路

可避开上海市限行

Navigation✓

Remaining tasks

复合任务的独特挑战

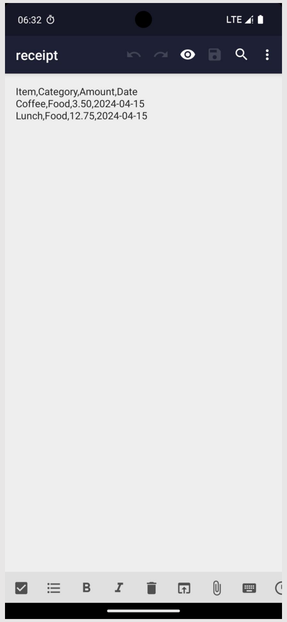
失败案例：信息传递失败，导致后续任务胡乱执行

User

[Instruction]: In Markor, open "receipt.md" and read the transactions listed in CSV format. Then add each transaction as a new expense in the arduia pro expense app.

M3A

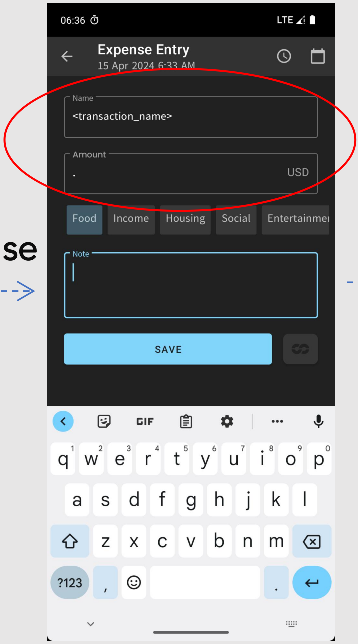
Open the receipt file



receipt

```
Item,Category,Amount,Date
Coffee,Food,3.50,2024-04-15
Lunch,Food,12.75,2024-04-15
```

Go to pro expense



Expense Entry
15 Apr 2024 6:33 AM

Name
<transaction_name>

Amount
. USD

Note

Confirm creation

Fail to Extract Proper Info ×

Type Faulty Info ×

复合任务的独特挑战

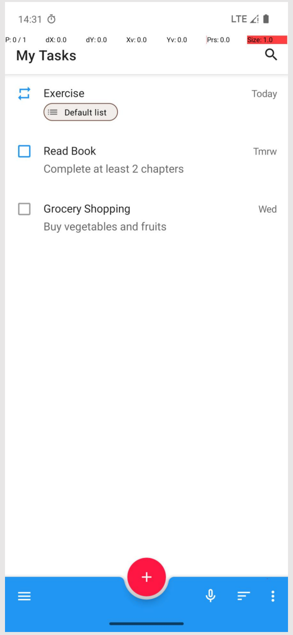
失败案例：**进度管理失败**，导致在不同场景间反复横跳

User

[Instruction]: In the Tasks app, create and save a new task named 'Exercise' repeating every day. Then open the Broccoli recipe app and delete the 'French Fries' recipe.

UI-TARS-7B-SFT

Create the task in Tasks app



14:31 LTE

P: 0 / 1 dX: 0.0 dY: 0.0 Xv: 0.0 Yv: 0.0 Prs: 0.0 Size: 1.0

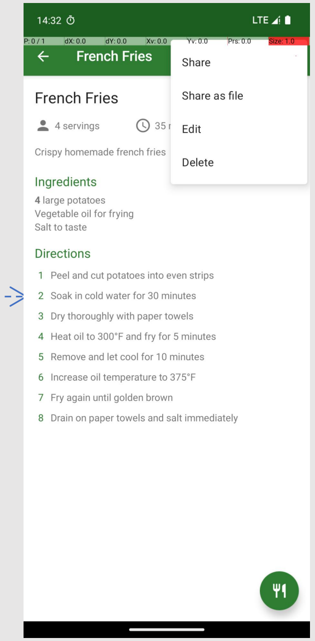
My Tasks

Exercise — Today
Default list

□ Read Book — Tmrw
Complete at least 2 chapters

□ Grocery Shopping — Wed
Buy vegetables and fruits

Go to Broccoli



14:32 LTE

P: 0 / 1 dX: 0.0 dY: 0.0 Xv: 0.0 Yv: 0.0 Prs: 0.0 Size: 1.0

← French Fries

Share
Share as file
Edit
Delete

French Fries

4 servings 35 r

Crispy homemade french fries

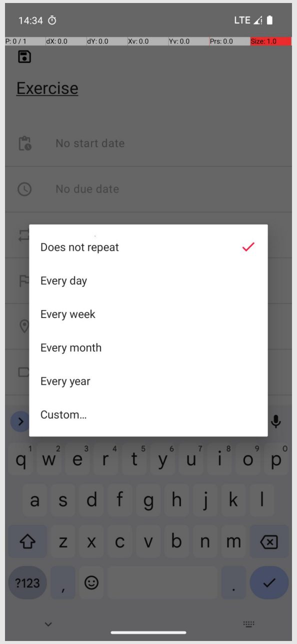
Ingredients

4 large potatoes
Vegetable oil for frying
Salt to taste

Directions

1 Peel and cut potatoes into even strips
2 Soak in cold water for 30 minutes
3 Dry thoroughly with paper towels
4 Heat oil to 300°F and fry for 5 minutes
5 Remove and let cool for 10 minutes
6 Increase oil temperature to 375°F
7 Fry again until golden brown
8 Drain on paper towels and salt immediately

Go to Tasks



14:34 LTE

P: 0 / 1 dX: 0.0 dY: 0.0 Xv: 0.0 Yv: 0.0 Prs: 0.0 Size: 1.0

Exercise

No start date

No due date

Does not repeat ✓
Every day
Every week
Every month
Every year
Custom...

Create the task

Delete the recipe

Creating the same task again

工作概览

1. 如何定义复合任务？

依据子任务依赖关系，定义三类复合任务指令，构造指令模板

2. 研究平台与基建

基于安卓搭建平台基建，支持任务环境自定义初始化和异构智能体可插拔适配

3. 系统实验与分析

50个中文&英文，在线&本地App，5大应用场景，5个工作流&专有模型智能体基线

全面测试揭示性能短板，分析实验揭示泛化困境

4. 高效解决方案

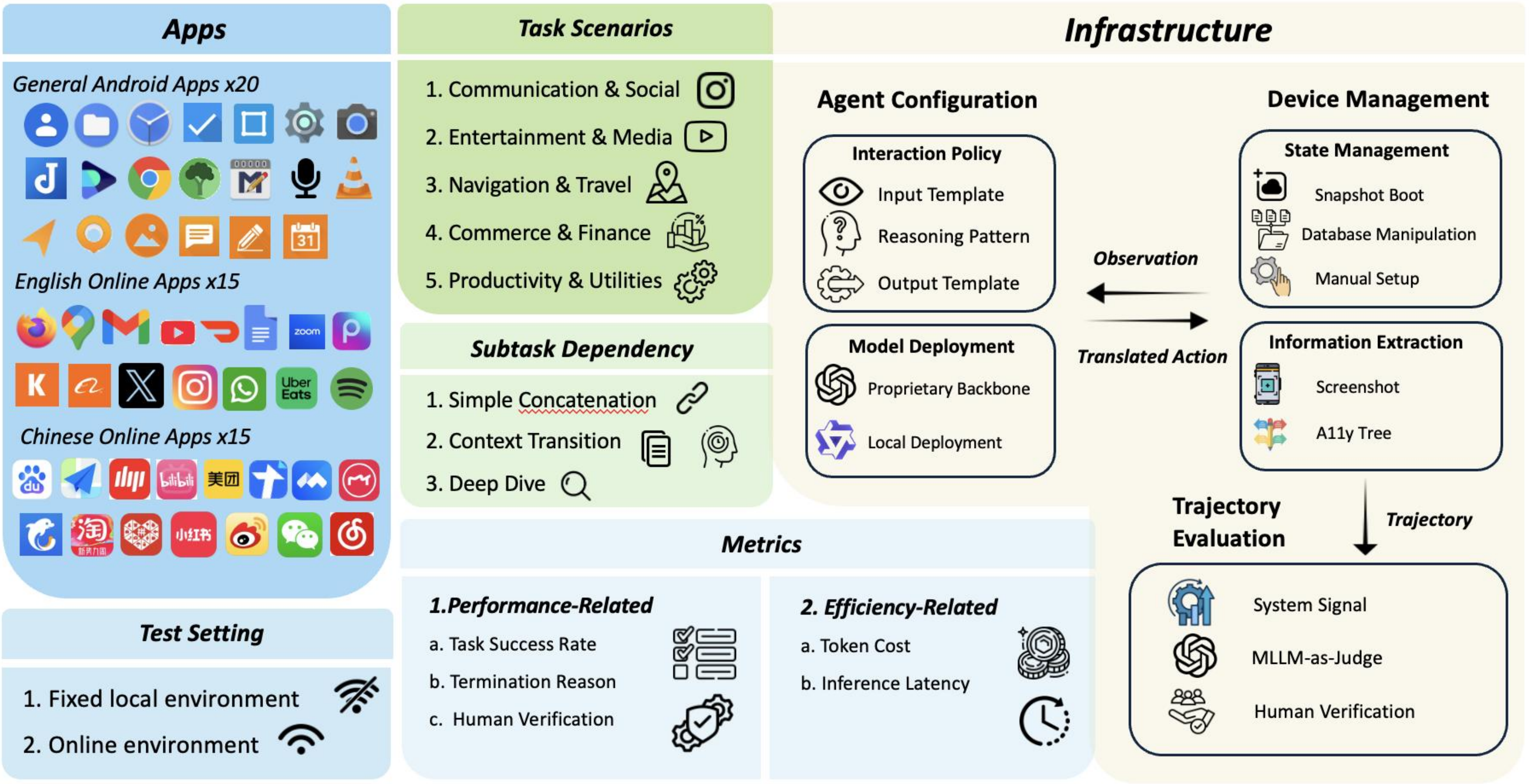
系统调度，语境收束，经济高效地提升复合任务成功率20%+

02

UI-NEXUS测试基准

UI-NEXUS测试基准概览

基于安卓平台，科学全面的OS智能体复合任务测试基准



Apps

General Android Apps x20

English Online Apps x15

Chinese Online Apps x15

Test Setting

1. Fixed local environment
2. Online environment

Task Scenarios

1. Communication & Social
2. Entertainment & Media
3. Navigation & Travel
4. Commerce & Finance
5. Productivity & Utilities

Subtask Dependency

1. Simple Concatenation
2. Context Transition
3. Deep Dive

Metrics

1.Performance-Related

a. Task Success Rate

b. Termination Reason

c. Human Verification

2. Efficiency-Related

a. Token Cost

b. Inference Latency

Infrastructure

Agent Configuration

Interaction Policy
- Input Template
- Reasoning Pattern
- Output Template

Model Deployment
- Proprietary Backbone
- Local Deployment

Observation

Translated Action

Device Management

State Management
- Snapshot Boot
- Database Manipulation
- Manual Setup

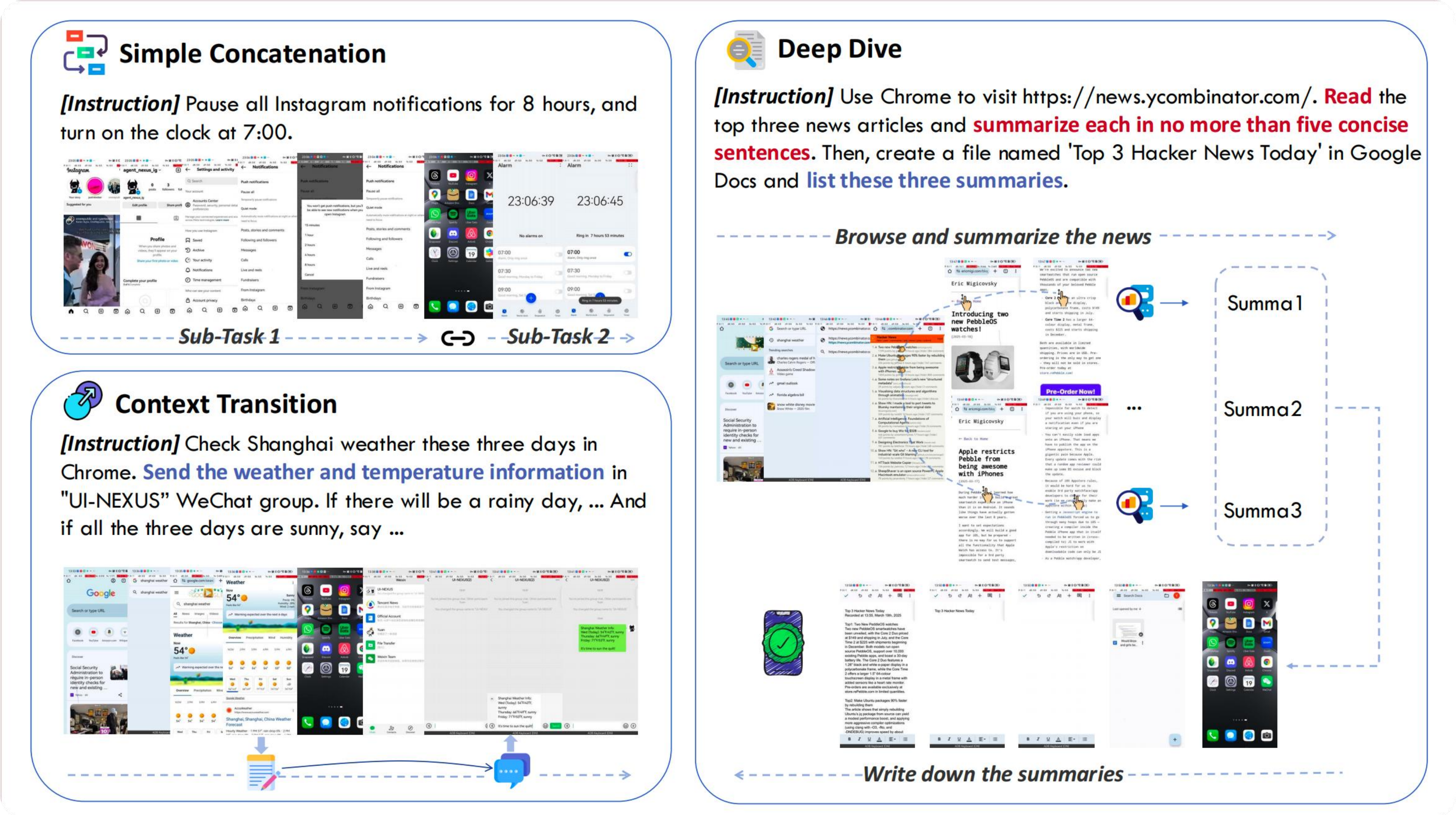
Information Extraction
- Screenshot
- A11y Tree

Trajectory

Trajectory Evaluation
- System Signal
- MLLM-as-Judge
- Human Verification

基于子任务依赖结构的复合指令分类

- Simple Concatenation (无关组合型）：无依赖的指令的直接组合
- Context Transition （语境传递型）：某些子任务有赖其他子任务的语境来实例化
- Deep Dive （深度分析型）：前一类的特殊情况：包含对中间语境信息的深度分析推理



Simple Concatenation

[Instruction] Pause all Instagram notifications for 8 hours, and turn on the clock at 7:00.

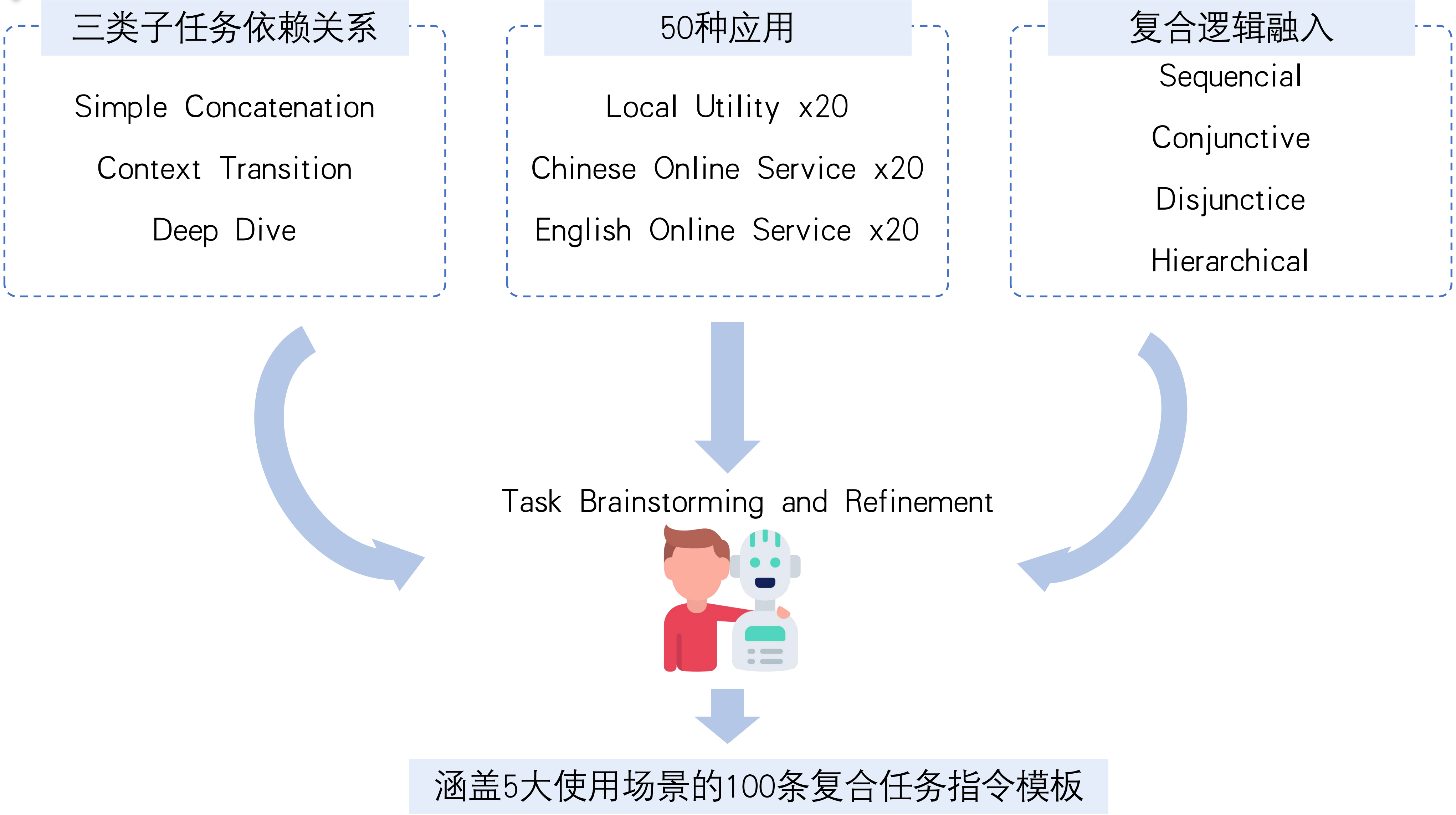
Context Transition

[Instruction] Check Shanghai weather these three days in Chrome. **Send the weather and temperature information** in "UI-NEXUS" WeChat group. If there will be a rainy day, ... And if all the three days are sunny, say ...

Deep Dive

[Instruction] Use Chrome to visit https://news.ycombinator.com/. **Read** the top three news articles and **summarize each in no more than five concise sentences**. Then, create a file named 'Top 3 Hacker News Today' in Google Docs and **list these three summaries**.

任务指令构造



三类子任务依赖关系

Simple Concatenation

Context Transition

Deep Dive

50种应用

Local Utility x20

Chinese Online Service x20

English Online Service x20

复合逻辑融入

Sequencial

Conjunctive

Disjunctice

Hierarchical

Task Brainstorming and Refinement

涵盖5大使用场景的100条复合任务指令模板

开发平台搭建

- 设备管理：构建定制化的测试环境

对于Pro Expense, Retro Music, Markor, Simple SMS Messenger等本地应用，采取ADB设定（如短信、蓝牙、WiFi状态）、数据库操作（如Pro Expense里的账单条目）、文件系统操作（如Markor笔记、文件管理器）结合的方式，实现根据本地的配置文件实现模拟器状态的初始化，构建统一的、可扩展的测试环境

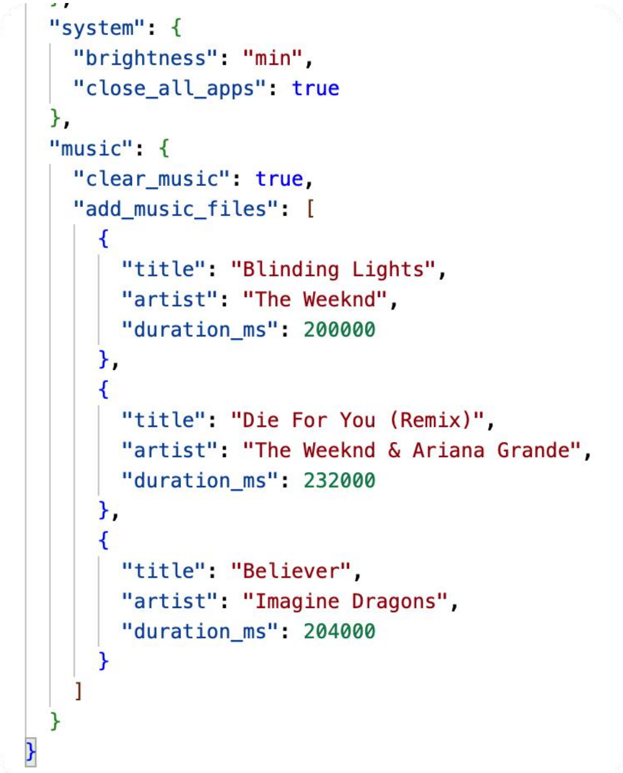
对于小红书、Instagram等在线服务应用，难以直接控制状态，用手动初始化确保测试准确

开发平台搭建

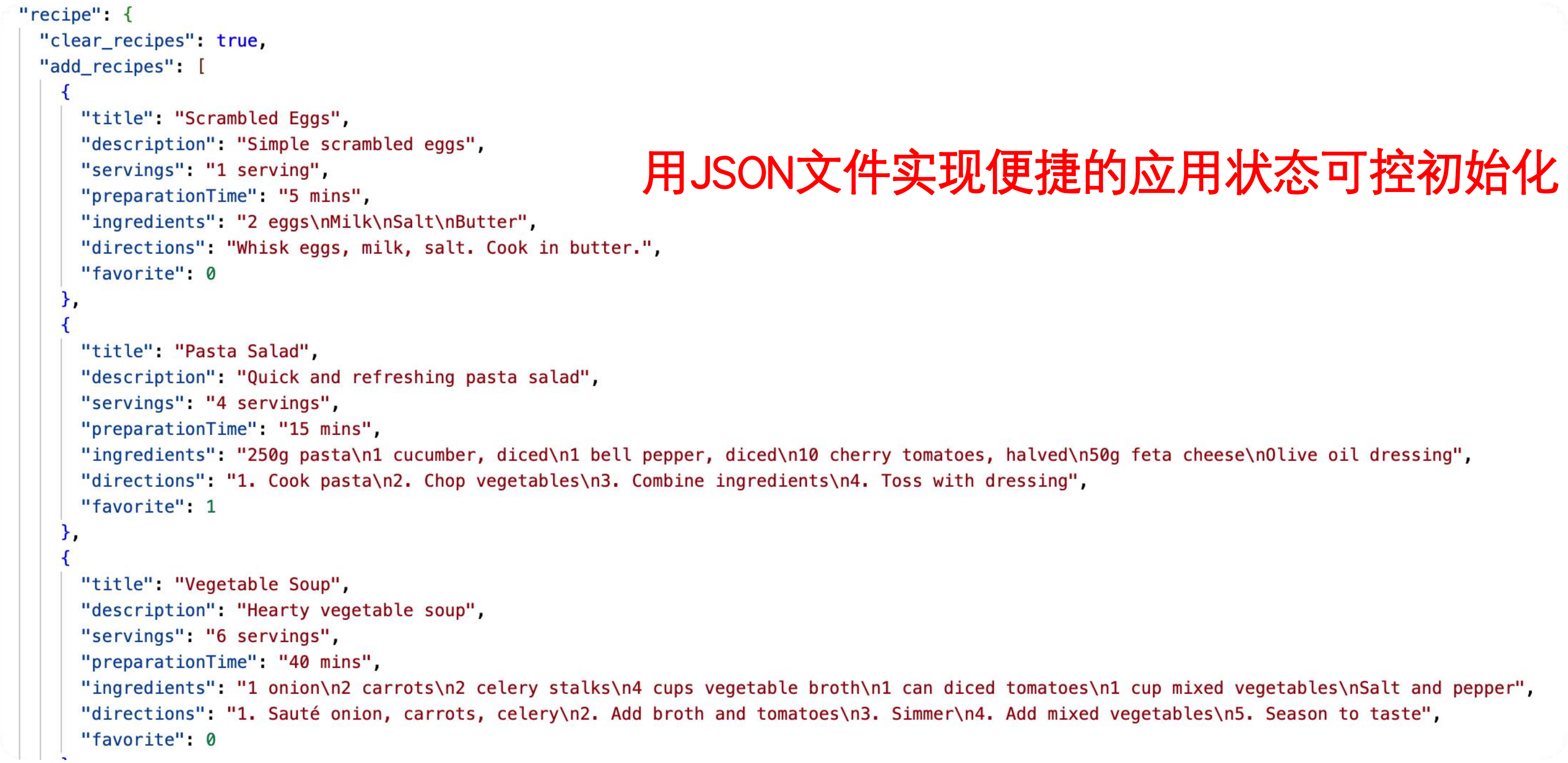
- 设备管理：构建定制化的测试环境

对于Pro Expense, Retro Music, Markor, Simple SMS Messenger等本地应用，采取ADB设定（如短信、蓝牙、WiFi状态）、数据库操作（如Pro Expense里的账单条目）、文件系统操作（如Markor笔记、文件管理器）结合的方式，实现根据本地的配置文件实现模拟器状态的初始化，构建统一的、可扩展的测试环境

对于小红书、Instagram等在线服务应用，难以直接控制状态，用手动初始化确保测试准确



```
  "system": {
    "brightness": "min",
    "close_all_apps": true
  },
  "music": {
    "clear_music": true,
    "add_music_files": [
      {
        "title": "Blinding Lights",
        "artist": "The Weeknd",
        "duration_ms": 200000
      },
      {
        "title": "Die For You (Remix)",
        "artist": "The Weeknd & Ariana Grande",
        "duration_ms": 232000
      },
      {
        "title": "Believer",
        "artist": "Imagine Dragons",
        "duration_ms": 204000
      }
    ]
  }
}
```



```
"recipe": {
  "clear_recipes": true,
  "add_recipes": [
    {
      "title": "Scrambled Eggs",
      "description": "Simple scrambled eggs",
      "servings": "1 serving",
      "preparationTime": "5 mins",
      "ingredients": "2 eggs\nMilk\nSalt\nButter",
      "directions": "Whisk eggs, milk, salt. Cook in butter.",
      "favorite": 0
    },
    {
      "title": "Pasta Salad",
      "description": "Quick and refreshing pasta salad",
      "servings": "4 servings",
      "preparationTime": "15 mins",
      "ingredients": "250g pasta\n1 cucumber, diced\n1 bell pepper, diced\n10 cherry tomatoes, halved\n50g feta cheese\nOlive oil dressing",
      "directions": "1. Cook pasta\n2. Chop vegetables\n3. Combine ingredients\n4. Toss with dressing",
      "favorite": 1
    },
    {
      "title": "Vegetable Soup",
      "description": "Hearty vegetable soup",
      "servings": "6 servings",
      "preparationTime": "40 mins",
      "ingredients": "1 onion\n2 carrots\n2 celery stalks\n4 cups vegetable broth\n1 can diced tomatoes\n1 cup mixed vegetables\nSalt and pepper",
      "directions": "1. Sauté onion, carrots, celery\n2. Add broth and tomatoes\n3. Simmer\n4. Add mixed vegetables\n5. Season to taste",
      "favorite": 0
```

用JSON文件实现便捷的应用状态可控初始化

开发平台搭建

- 设备管理：构建定制化的测试环境

对于Pro Expense, Retro Music, Markor, Simple SMS Messenger等本地应用，采取ADB设定（如短信、蓝牙、WiFi状态）、数据库操作（如Pro Expense里的账单条目）、文件系统操作（如Markor笔记、文件管理器）结合的方式，实现根据本地的配置文件实现模拟器状态的初始化，构建统一的、可扩展的测试环境

对于小红书、Instagram等在线服务应用，难以直接控制状态，用手动初始化确保测试准确

- 智能体配置：集成主流手机智能体框架

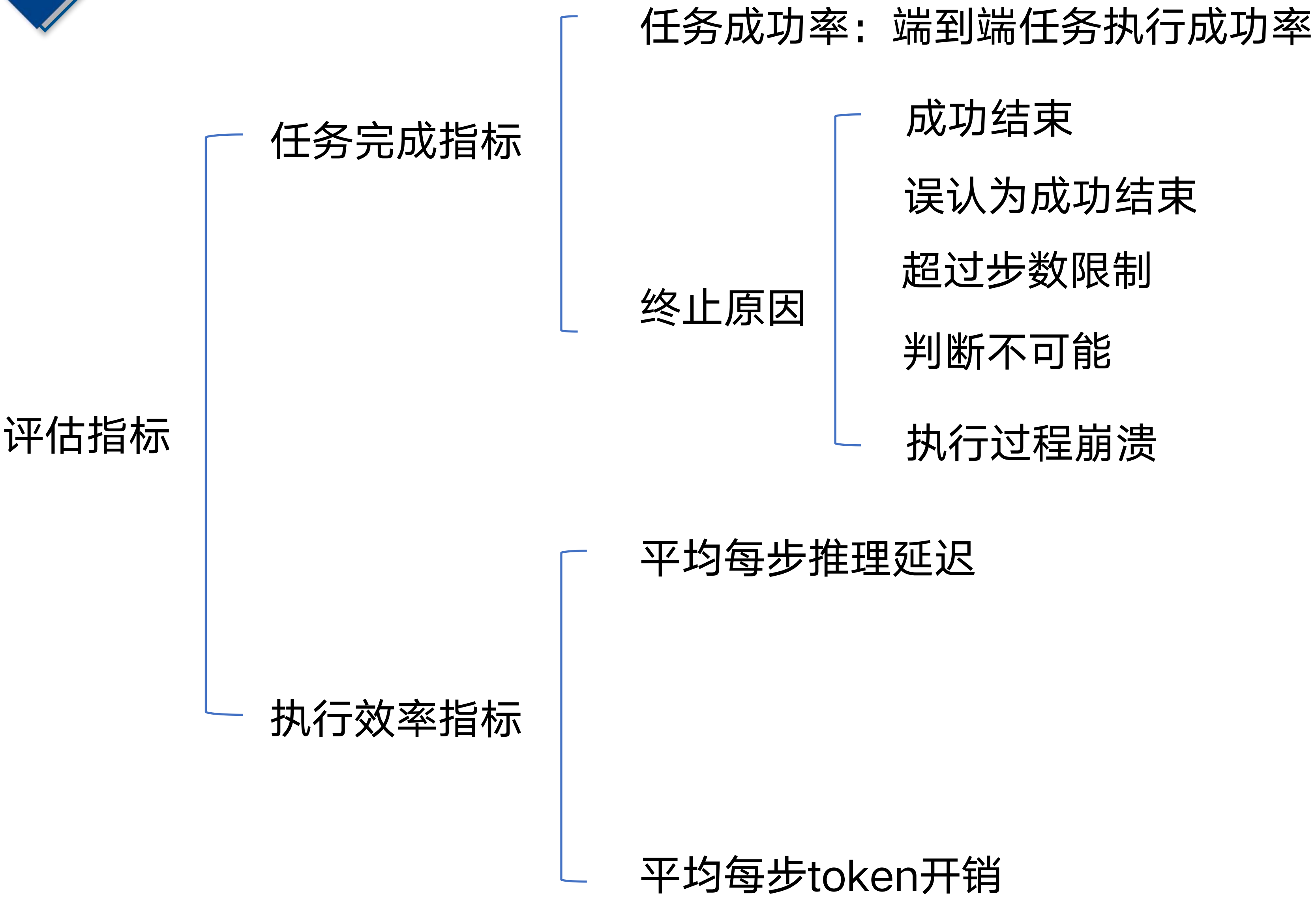
集成包括Mobile-Agent-E和M3A等Agentic Workflow和UI-TARS等Agent-as-a-Model的手机智能体框架，支持与模拟器/真机进行交互，并且记录完整的输入输出、截图轨迹、token消耗与延迟等

- 轨迹评估：评估任务完成情况

综合利用系统信号提取、大模型打分、人类验证判断轨迹成功

计算平均延迟、平均每步开销等指标

评估指标



- 评估指标
  - 任务完成指标
    - 任务成功率：端到端任务执行成功率
    - 终止原因
      - 成功结束
      - 误认为成功结束
      - 超过步数限制
      - 判断不可能
      - 执行过程崩溃
  - 执行效率指标
    - 平均每步推理延迟
    - 平均每步token开销

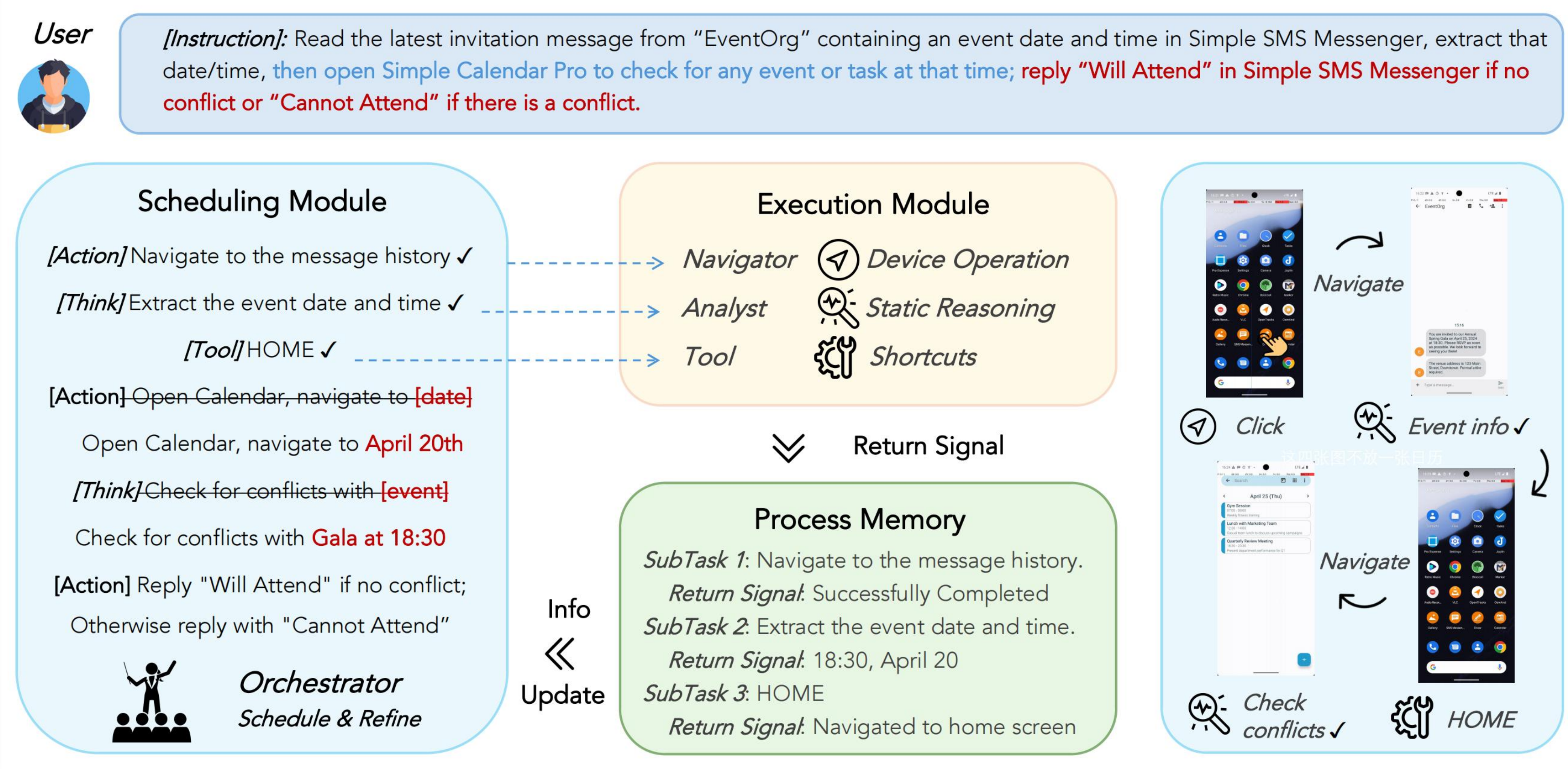
03

Agent-NEXUS调度系统

面向复杂长程场景的任务调度系统

单个智能体模型难以处理多场景协调和复杂依赖，容易出现语境溢出、进度混乱问题

构建智能体**任务调度系统**，对复杂任务进行拆解和调度



User

[Instruction]: Read the latest invitation message from "EventOrg" containing an event date and time in Simple SMS Messenger, extract that date/time, then open Simple Calendar Pro to check for any event or task at that time; reply "Will Attend" in Simple SMS Messenger if no conflict or "Cannot Attend" if there is a conflict.

Scheduling Module

[Action] Navigate to the message history ✓

[Think] Extract the event date and time ✓

[Tool] HOME ✓

[Action] ~~Open Calendar, navigate to [date]~~

Open Calendar, navigate to April 20th

[Think] ~~Check for conflicts with [event]~~

Check for conflicts with Gala at 18:30

[Action] Reply "Will Attend" if no conflict;

Otherwise reply with "Cannot Attend"

Orchestrator

Schedule & Refine

Execution Module

Navigator — *Device Operation*

Analyst — *Static Reasoning*

Tool — *Shortcuts*

Return Signal

Info

Update

Process Memory

SubTask 1: Navigate to the message history.

Return Signal: Successfully Completed

SubTask 2: Extract the event date and time.

Return Signal: 18:30, April 20

SubTask 3: HOME

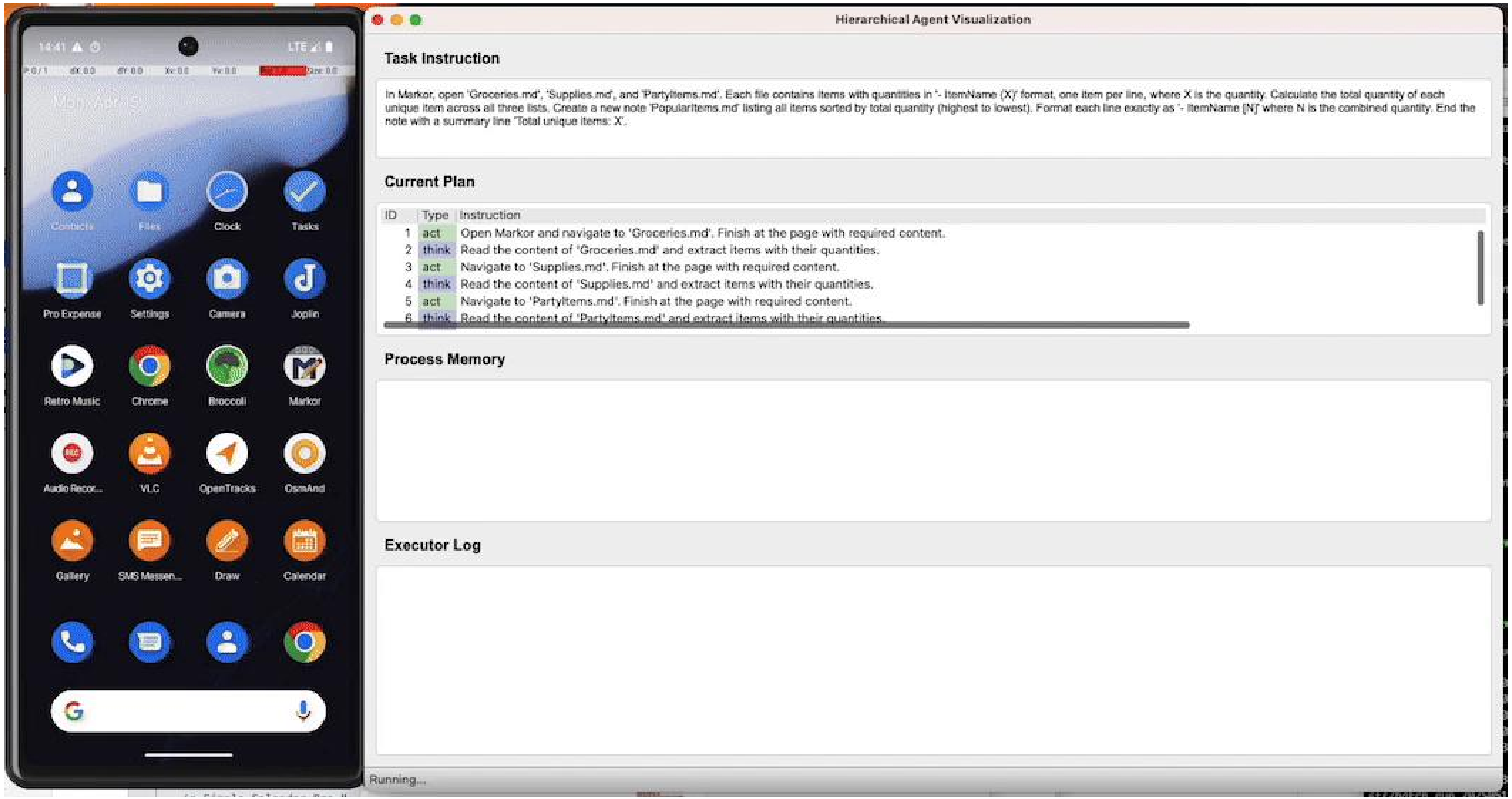
Return Signal: Navigated to home screen

面向复杂长程场景的任务调度系统

通过系统级别的调度和管理，实现信息的异步获取、传递和整合，突破已有架构局限

In Markor, open 'Groceries.md', 'Supplies.md', and 'PartyItems.md'. Each file contains items with quantities in '- ItemName (X)' format, one item per line, where X is the quantity. Calculate the total quantity of each unique item across all three lists. Create a new note 'PopularItems.md' listing all items sorted by total quantity (highest to lowest). Format each line exactly as '- ItemName (N)' where N is the combined quantity. End the note with a summary line 'Total unique items: X'.

【对三个购物清单进行合并和整理】分别读取三个文件的信息 -> 对内容整合、累加、分析 -> 将分析结果写入新文件



Task Instruction

In Markor, open 'Groceries.md', 'Supplies.md', and 'PartyItems.md'. Each file contains items with quantities in '- ItemName (X)' format, one item per line, where X is the quantity. Calculate the total quantity of each unique item across all three lists. Create a new note 'PopularItems.md' listing all items sorted by total quantity (highest to lowest). Format each line exactly as '- ItemName (N)' where N is the combined quantity. End the note with a summary line 'Total unique items: X'.

Current Plan

ID	Type	Instruction
1	act	Open Markor and navigate to 'Groceries.md'. Finish at the page with required content.
2	think	Read the content of 'Groceries.md' and extract items with their quantities.
3	act	Navigate to 'Supplies.md'. Finish at the page with required content.
4	think	Read the content of 'Supplies.md' and extract items with their quantities.
5	act	Navigate to 'PartyItems.md'. Finish at the page with required content.
6	think	Read the content of 'PartyItems.md' and extract items with their quantities.

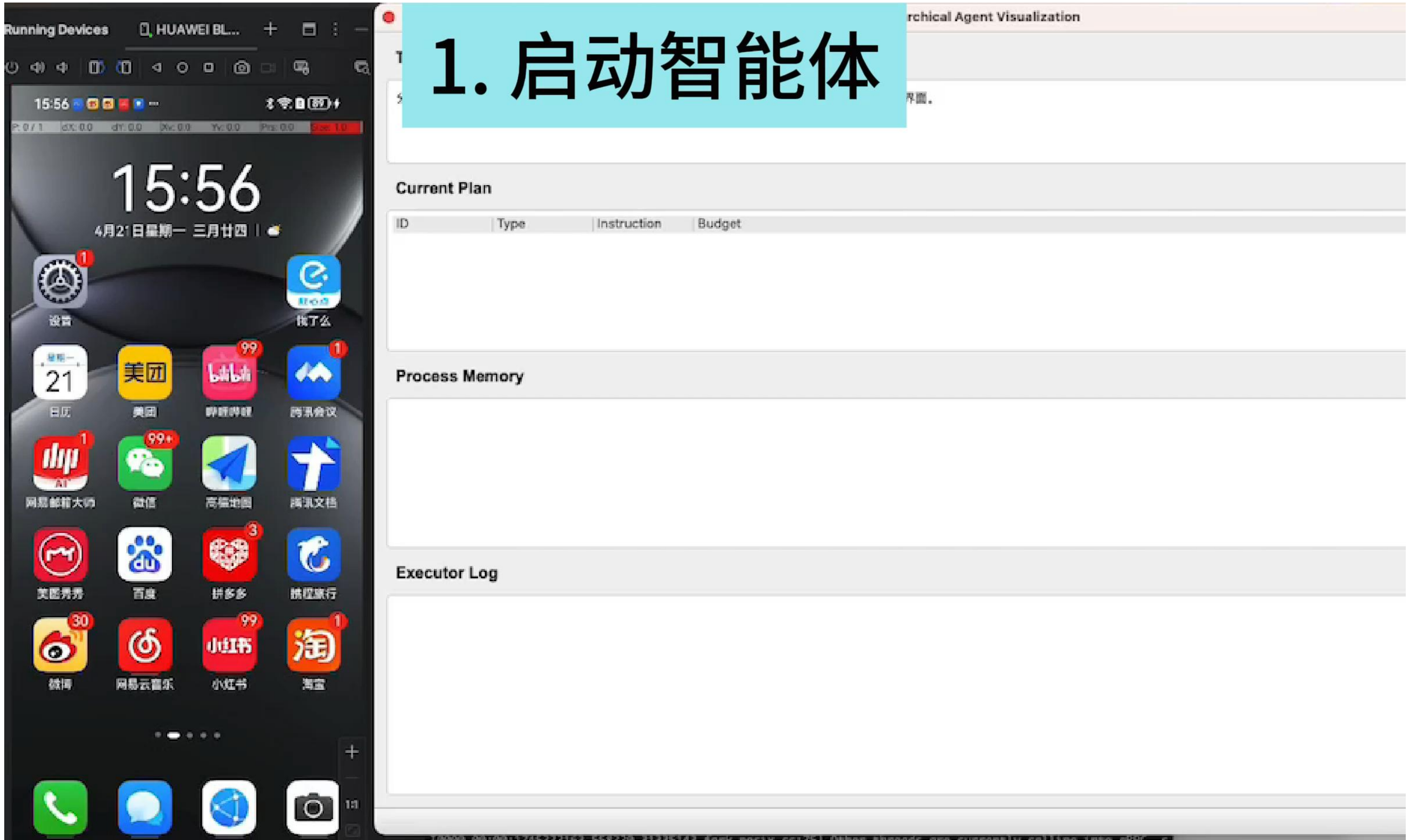
Process Memory

Executor Log

Running...

面向复杂长程场景的任务调度系统

指令：分别在**美团、饿了么**里搜索**肯德基超级全家桶**，然后在价格**最便宜的一个平台**下单，停留在下单界面。



1. 启动智能体

rchical Agent Visualization

Current Plan

ID | Type | Instruction | Budget

Process Memory

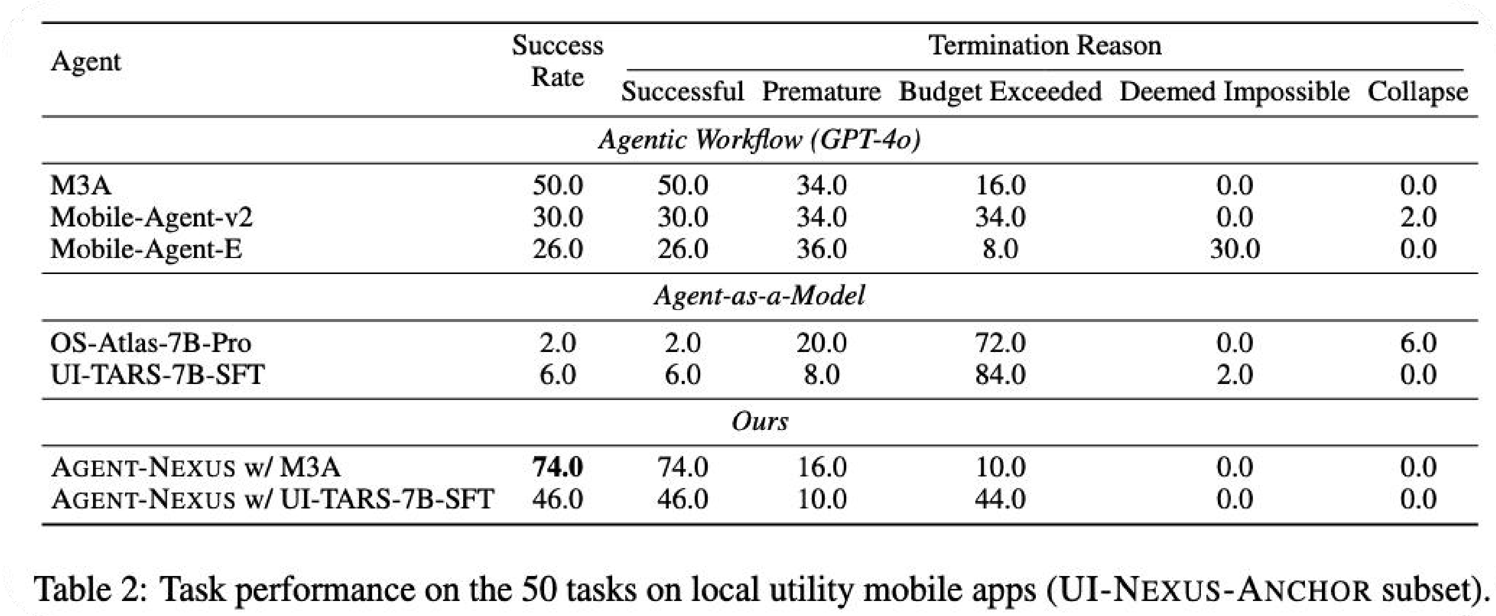
Executor Log

04

实验分析

任务完成情况测评

- 复合任务对现有智能体造成较大挑战，所有智能体在所有子集中任务完成率不超过50%
- 在线服务应用由于UI设计复杂、环境干扰多等，构成了更大的挑战
- 相比之下，基于GPT-4o的Agentic Workflow在处理复合任务时比Agent-as-a-Model更鲁棒
- Agent-NEXUS大幅度提升了智能体的任务完成率，尤其是对于UI-TARS-7B-SFT



Agent	Success Rate	Termination Reason				
		Successful	Premature	Budget Exceeded	Deemed Impossible	Collapse
Agentic Workflow (GPT-4o)						
M3A	50.0	50.0	34.0	16.0	0.0	0.0
Mobile-Agent-v2	30.0	30.0	34.0	34.0	0.0	2.0
Mobile-Agent-E	26.0	26.0	36.0	8.0	30.0	0.0
Agent-as-a-Model						
OS-Atlas-7B-Pro	2.0	2.0	20.0	72.0	0.0	6.0
UI-TARS-7B-SFT	6.0	6.0	8.0	84.0	2.0	0.0
Ours						
AGENT-NEXUS w/ M3A	**74.0**	74.0	16.0	10.0	0.0	0.0
AGENT-NEXUS w/ UI-TARS-7B-SFT	46.0	46.0	10.0	44.0	0.0	0.0

Table 2: Task performance on the 50 tasks on local utility mobile apps (UI-NEXUS-ANCHOR subset).

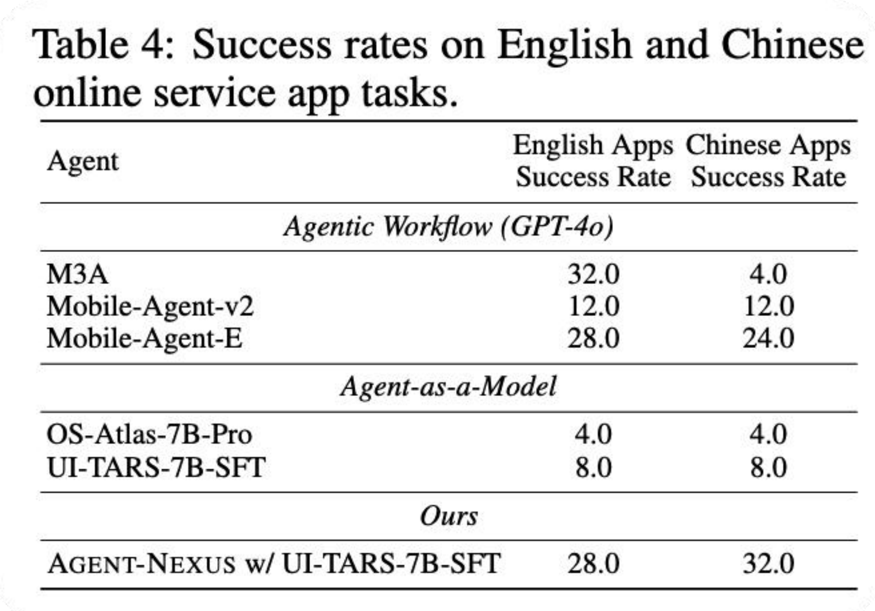


Table 4: Success rates on English and Chinese online service app tasks.

Agent	English Apps Success Rate	Chinese Apps Success Rate
Agentic Workflow (GPT-4o)		
M3A	32.0	4.0
Mobile-Agent-v2	12.0	12.0
Mobile-Agent-E	28.0	24.0
Agent-as-a-Model		
OS-Atlas-7B-Pro	4.0	4.0
UI-TARS-7B-SFT	8.0	8.0
Ours		
AGENT-NEXUS w/ UI-TARS-7B-SFT	28.0	32.0

任务执行效率测评

- 基于GPT-4o的Agentic Workflow在处理复合任务时比Agent-as-a-Model更鲁棒，但是时间和token开销很大，距离实际部署应用尚有差距，在每步都采用多智能体协同决策带来较大的计算冗余
- Agent-as-a-Model有显著更加轻便高效，且易于利用领域知识个性化强化等优势，但是面对复合任务较容易崩溃

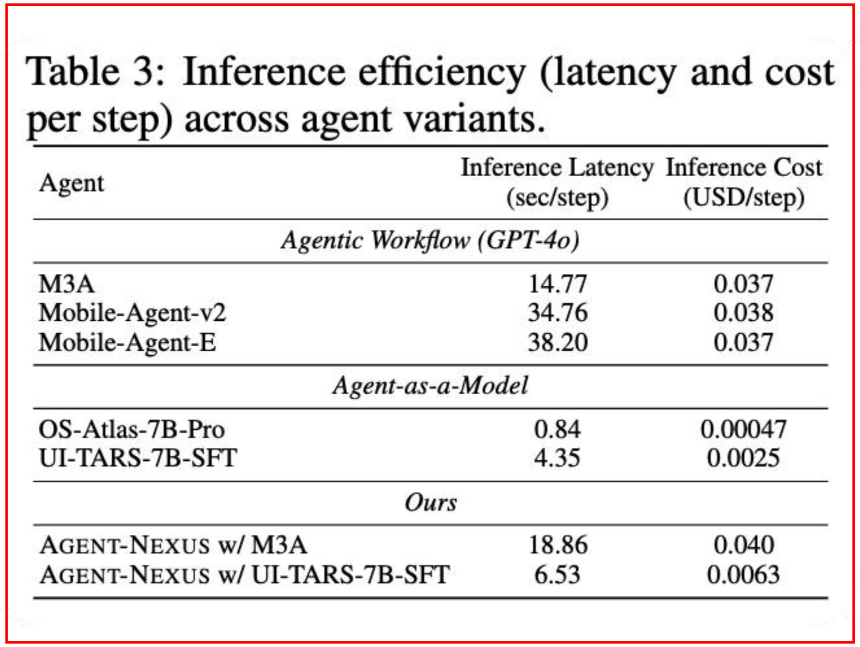
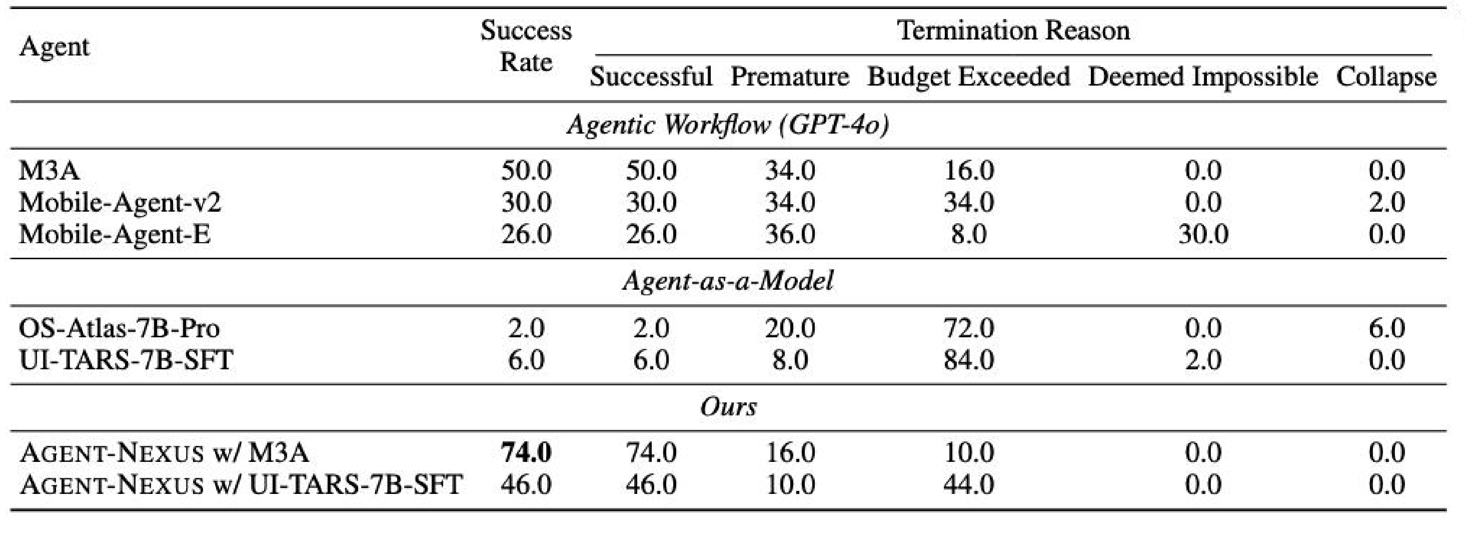


Table 3: Inference efficiency (latency and cost per step) across agent variants.

Agent	Inference Latency (sec/step)	Inference Cost (USD/step)
Agentic Workflow (GPT-4o)		
M3A	14.77	0.037
Mobile-Agent-v2	34.76	0.038
Mobile-Agent-E	38.20	0.037
Agent-as-a-Model		
OS-Atlas-7B-Pro	0.84	0.00047
UI-TARS-7B-SFT	4.35	0.0025
Ours		
AGENT-NEXUS w/ M3A	18.86	0.040
AGENT-NEXUS w/ UI-TARS-7B-SFT	6.53	0.0063

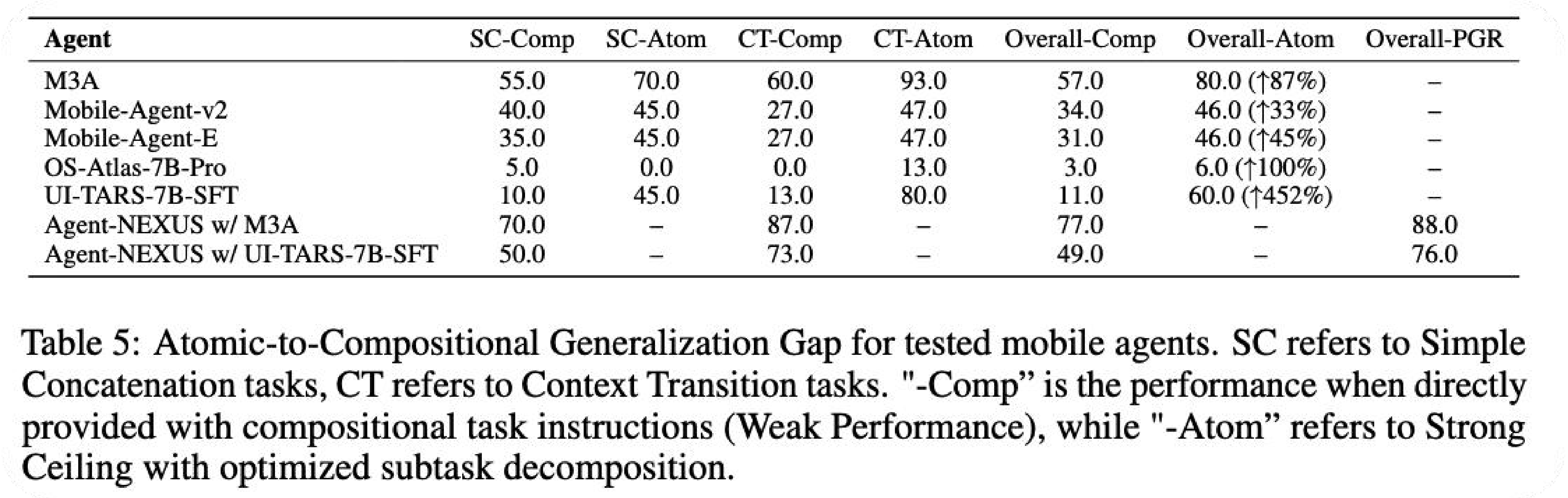


Agent	Success Rate	Termination Reason				
		Successful	Premature	Budget Exceeded	Deemed Impossible	Collapse
Agentic Workflow (GPT-4o)						
M3A	50.0	50.0	34.0	16.0	0.0	0.0
Mobile-Agent-v2	30.0	30.0	34.0	34.0	0.0	2.0
Mobile-Agent-E	26.0	26.0	36.0	8.0	30.0	0.0
Agent-as-a-Model						
OS-Atlas-7B-Pro	2.0	2.0	20.0	72.0	0.0	6.0
UI-TARS-7B-SFT	6.0	6.0	8.0	84.0	2.0	0.0
Ours						
AGENT-NEXUS w/ M3A	**74.0**	74.0	16.0	10.0	0.0	0.0
AGENT-NEXUS w/ UI-TARS-7B-SFT	46.0	46.0	10.0	44.0	0.0	0.0

Table 2: Task performance on the 50 tasks on local utility mobile apps (UI-NEXUS-ANCHOR subset).

分析实验：原子到复合能力泛化

- 选择35个Simple Concatenation和Context Transition类型任务
- 分别测试：（i）直接给定复合指令（ii）分别给最优的手动原子指令拆分（iii）复合指令+调度系统
- 各智能体都呈现显著的原子-复合泛化损失，其中UI-TARS-7B-SFT尤为显著
- Agent-NEXUS通过任务调度实现了语境收束，逼近手动拆分的最优表现

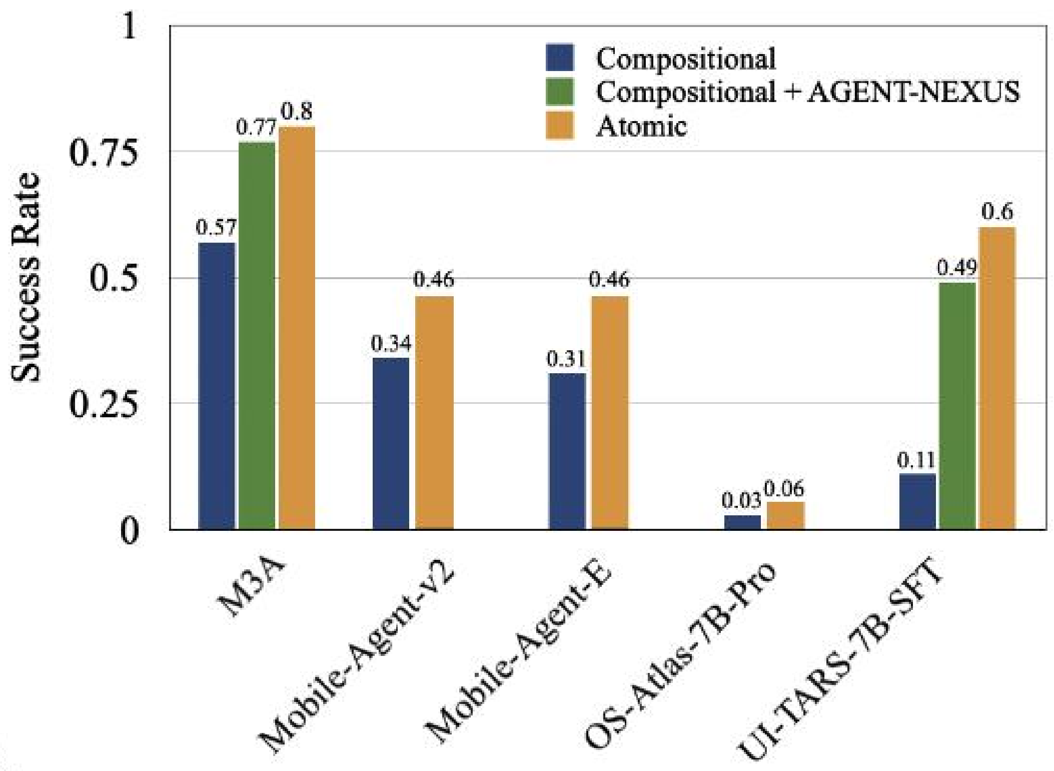


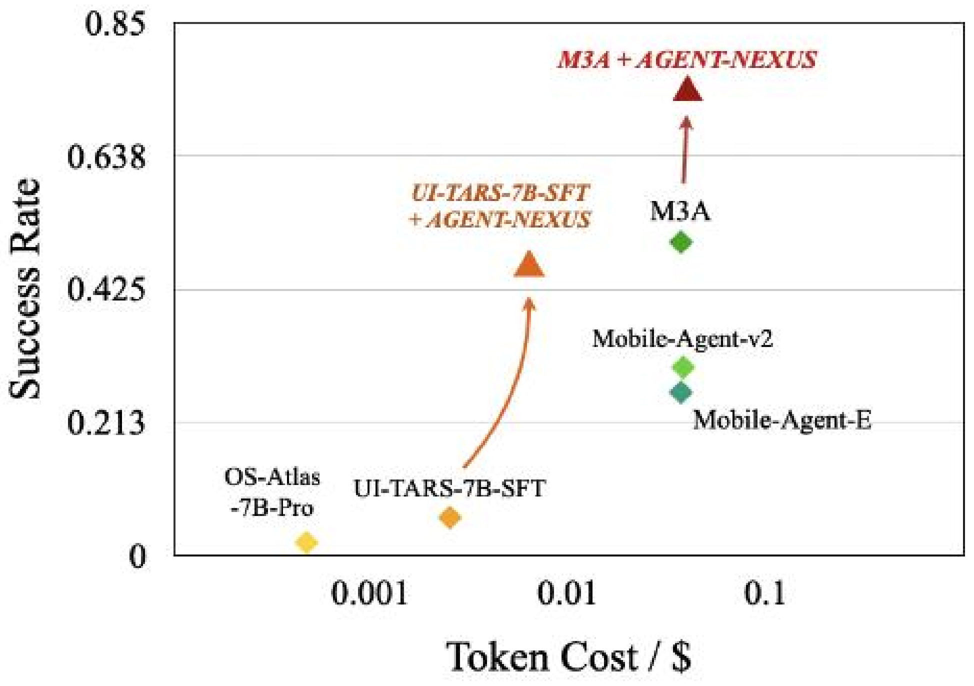
Agent	SC-Comp	SC-Atom	CT-Comp	CT-Atom	Overall-Comp	Overall-Atom	Overall-PGR
M3A	55.0	70.0	60.0	93.0	57.0	80.0 (↑87%)	–
Mobile-Agent-v2	40.0	45.0	27.0	47.0	34.0	46.0 (↑33%)	–
Mobile-Agent-E	35.0	45.0	27.0	47.0	31.0	46.0 (↑45%)	–
OS-Atlas-7B-Pro	5.0	0.0	0.0	13.0	3.0	6.0 (↑100%)	–
UI-TARS-7B-SFT	10.0	45.0	13.0	80.0	11.0	60.0 (↑452%)	–
Agent-NEXUS w/ M3A	70.0	–	87.0	–	77.0	–	88.0
Agent-NEXUS w/ UI-TARS-7B-SFT	50.0	–	73.0	–	49.0	–	76.0

Table 5: Atomic-to-Compositional Generalization Gap for tested mobile agents. SC refers to Simple Concatenation tasks, CT refers to Context Transition tasks. "-Comp” is the performance when directly provided with compositional task instructions (Weak Performance), while "-Atom” refers to Strong Ceiling with optimized subtask decomposition.

实验结果可视化

- 各智能体都呈现显著的原子-复合泛化损失，Agent-NEXUS调度系统显著地弥补了gap
- 通过将高阶调度和低阶执行解耦，Agent-NEXUS在开销增加可控的同时大幅提升完成率





05

未来展望

未来展望

- 基于强化学习的长程任务规划调度能力增强

- 更加精细的调度方式，如子任务的并行

- 多平台、跨平台任务

- 更加多元的多智能体协同架构和复杂长程任务

郭源

敬请批评指正！

▲ 郭源

▲ 2025.5.29