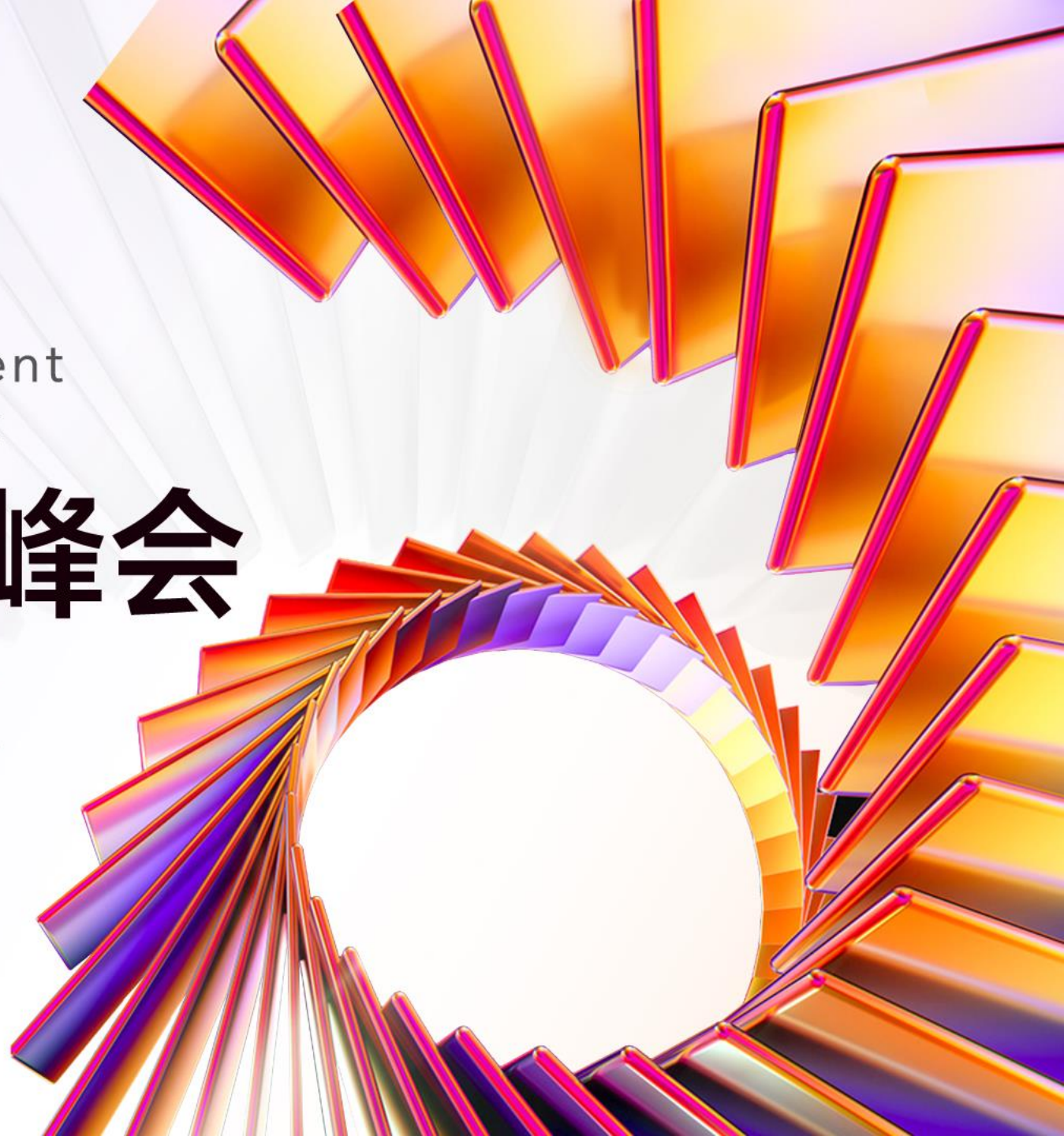
AiDD 8th 2025

第8届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发

11月14-15日 | 深圳



AiDD × K

2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会

Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞家与需求发现

论坛2: AI原生产品战路与商业模式重构

论坛3: AgenticAI产品创新与交互设计

Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

2-hour Speech: 回归本质

用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，仍然考验“看见”的“同理心”

Panel 1: 出海前瞻

“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式

Panel 2: 失败复盘

为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长

Reinventing Products with Intelligence, Driven by Data

80+ 业界大咖 汇聚产品大咖的真知灼见

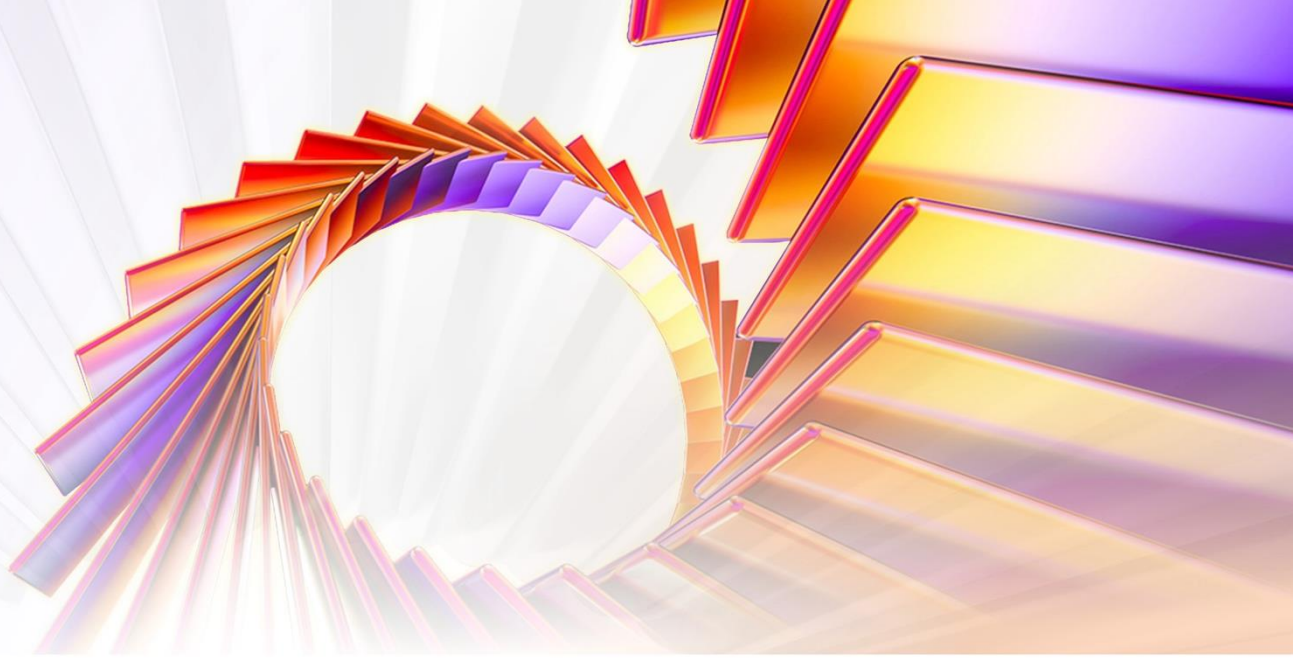
40+ 创新案例 开拓全新AI+产品视角

10+ 分论坛 涵盖产品到商业增长的全链路

800+ 听众规模 共同探索产品的智能重构

扫码查看会议详情

AiDD 8th 2025 | 11月14-15日 | 深圳
第8届 AI+ Development Digital Summit
AI+研发数字峰会
拥抱AI 重塑研发



Nebula-GUI Agent：精准快稳的端到端屏幕操作解决方案

张凯莉 | 中兴通讯股份有限公司

张凯莉

高级AI算法工程师

中兴通讯股份有限公司高级AI算法工程师，目前负责多模态大模型领域的技术研究和产品落地，包括GUI Agent、智能意图识别&FC、智慧家庭助手等。多次参加百度、python开发者大会等开源活动，github开源社区Adlik项目（深度学习推理优化加速工具链）的主要维护者之一。

目录
CONTENTS

I. 背景

II. 痛点

III. 整体方案

IV. 具体实现

V. 总结与展望

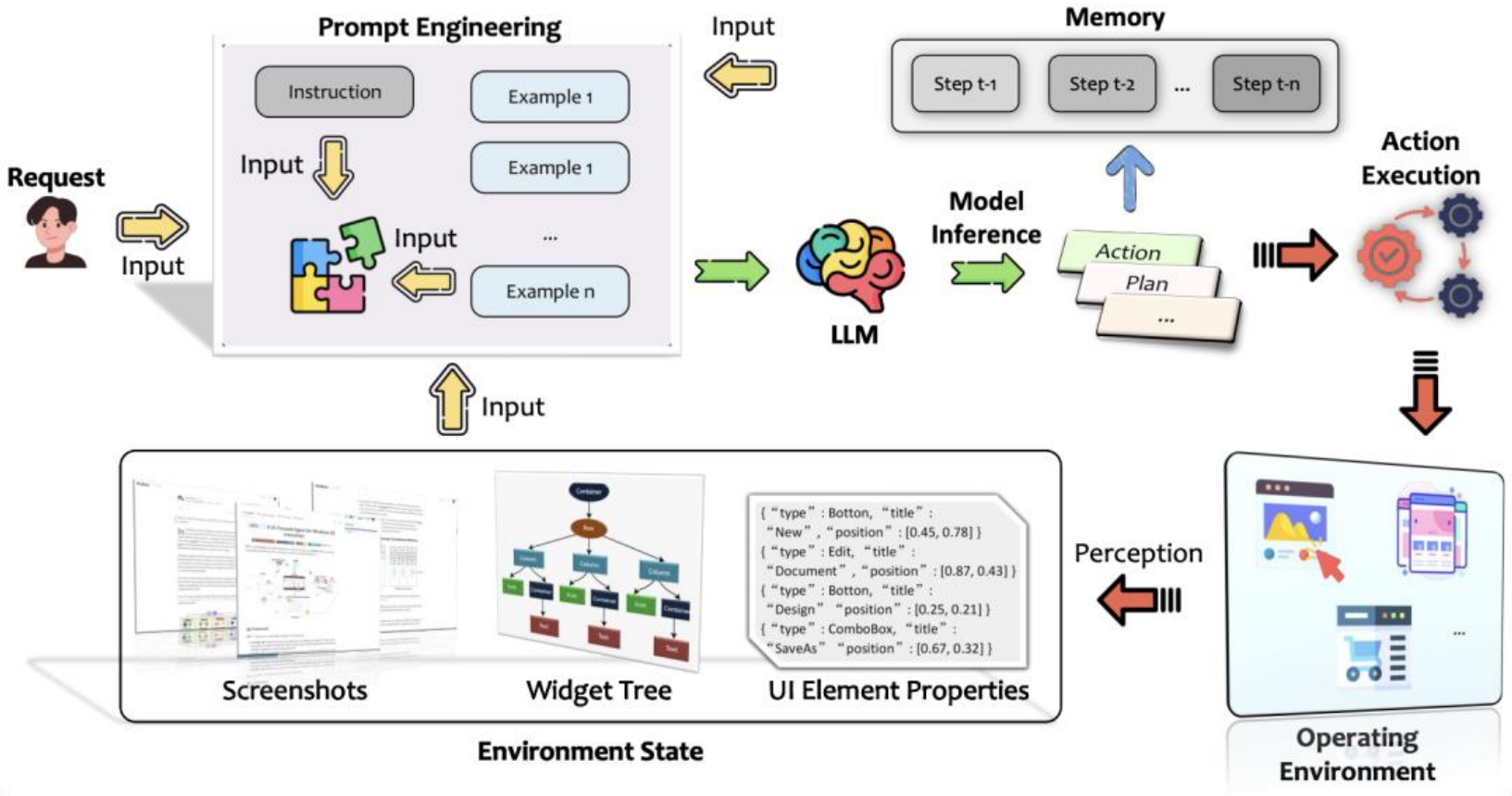
PART 01

背景

AI 驱动 GUI Agent 诞生

背景

- 大语言模型（LLM）和多模态大模型（VLM）的迅猛发展，推动AI向具备感知、推理、执行复杂任务的智能体演进
- GUI Agent 通过自然语言理解用户意图，能够像人类一样 "看屏幕、点按钮、做判断"，有望成为新的“超级入口”



GUI-Agent 的业界现状

- 业界现状

手机端到端操作GUI-Agent技术飞速发展，算法设计上快速迭代创新，国内主要科技公司相继发布相关模型或研究成果。

- 商用情况

25年6月30
支持购物、订票、下视频

25年7月30
支持点餐、订票、写好评

25年8月30
支持淘宝购物和抖音娱乐

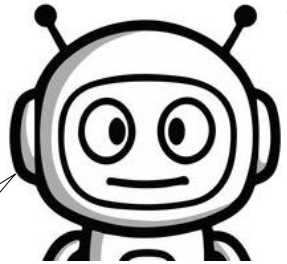
25年10月10
支持代打客服电话等

典型应用：购物比价

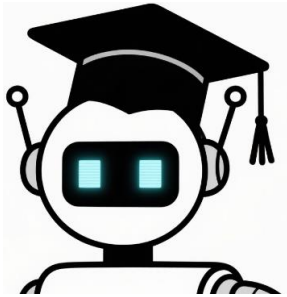
用户需求

帮我购买一台努比亚Z70 Ultra手机

请问您想买多大的内存?

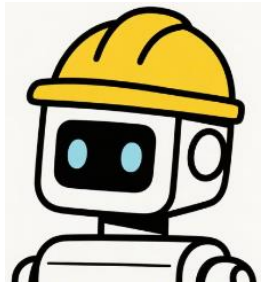


交互功能

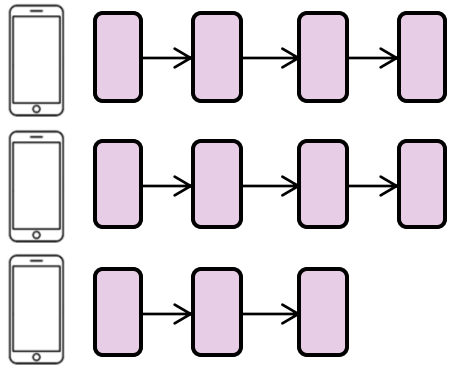


任务规划

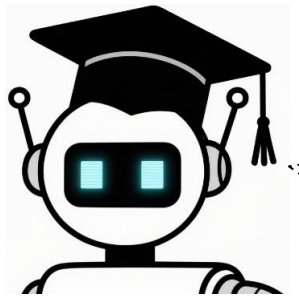
用户需要购买一台努比亚Z70 Ultra手机，我首先要调研一下各个购物平台上的该手机的价格，而常见的购物平台有京东，淘宝，拼多多。。。所以，用户任务拆解成一下一下几个子任务：
（1）分别在各个购物平台检索“努比亚Z70 Ultra手机”，并获取配置、价格、评价等。
（2）将上述检索结果进行总结反馈。



子任务执行



离屏渲染，后台并行执行子任务，不影响用户使用



任务总结

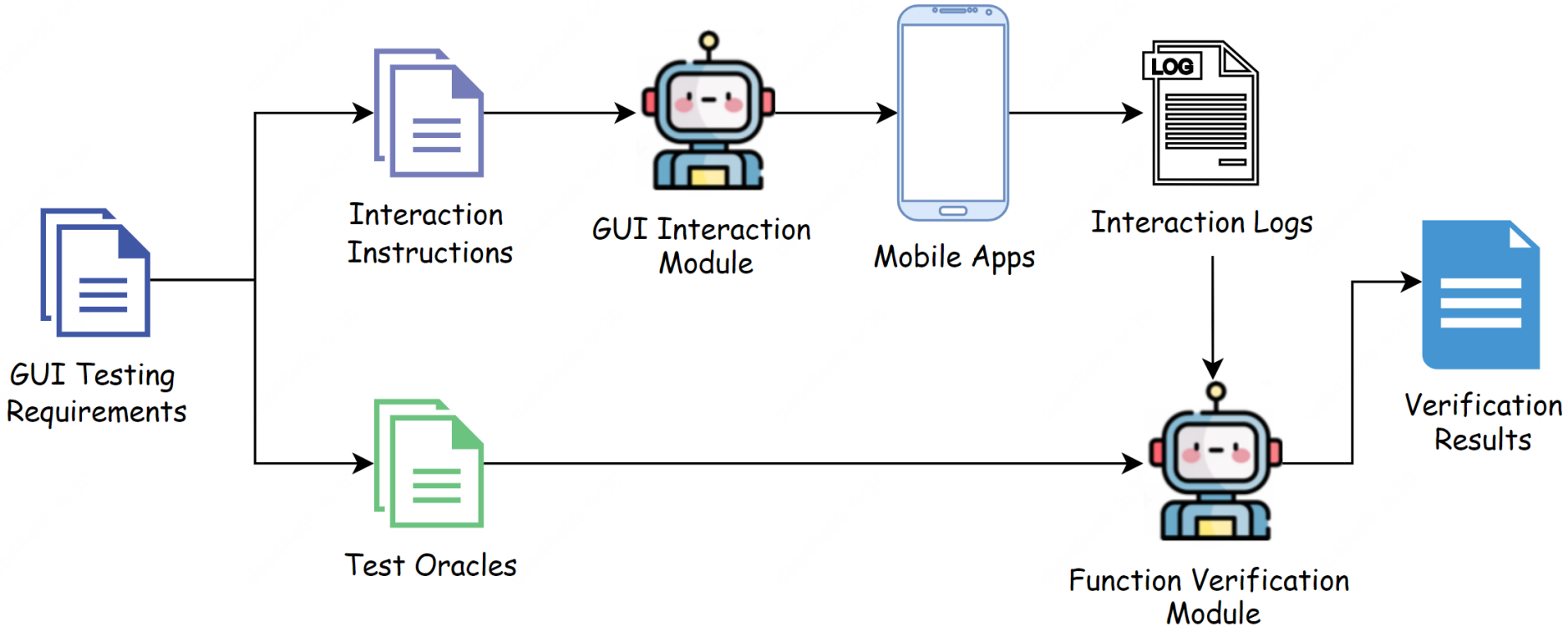
平台	原价	叠加优惠后到手价	优惠说明
京东自营	4599 元	3499-3849 元	满减券 + 国补 15 %（最高再减 500 元）
天猫官方旗舰店	4699 元	4099 元	店铺直降 600 元 + 天猫品类券
拼多多（百亿补贴）	4599 元	约 3799-3899 元	百亿补贴券包 + 限时秒杀（券量有限，需要抢）
抖音电商	4599 元	3999 元左右	直播间专属券/平台满减券

面向场景的Agent 服务：

- 一键比价
- 游戏代练
- 拼多多刷券
- 定时抢票
- 社交自动回
- 广告代看
- 更多个性定制...

典型应用：手机应用开发智能化测试

- 在快速迭代的软件开发中，UI元素和布局经常变动，这会导致大量基于坐标或固定元素定位的传统测试脚本失效
- GUI Agent能够感知环境变化（如按钮ID改变），以意图为目标使用自然语言描述进行测试，不但聚焦了应用是否能够实现客户意图，还显著降低了测试脚本的维护成本和测试人员的专业技术要求



《AUITestAgent: Automatic Requirements Oriented GUI Function Testing》

GUI-Agent 的技术趋势

快

云侧模型---->端侧模型

准

单体模型---->多Agent协同

泛

传统范式---->在线学习

精

通用场景---->垂域赋能

PART 02

痛点

GUI-Agent 的痛点问题

➢ 技术实现痛点

- 界面变更频繁
- 端侧资源受限
- 容错能力不足

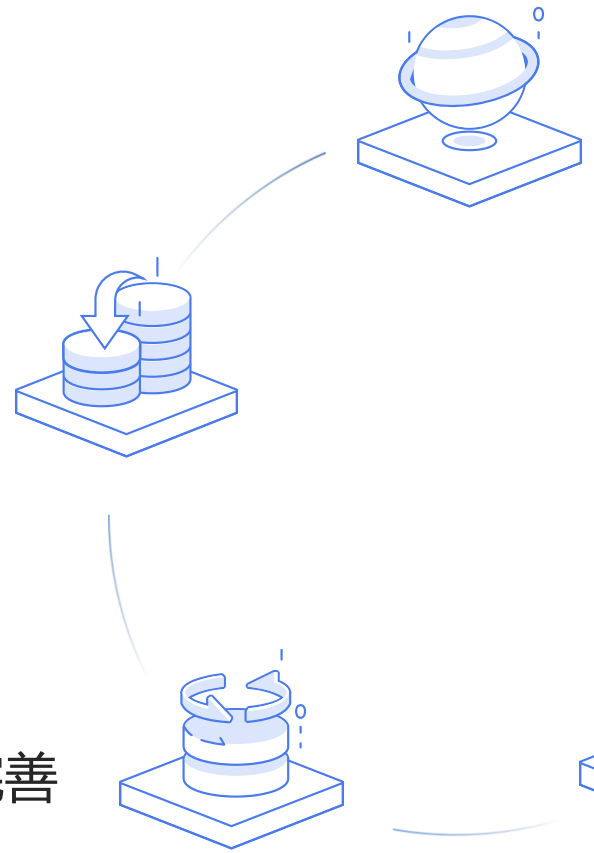
➢ 场景适配痛点

- 跨平台兼容性差
- 意图理解不合拍
- 垂直场景能力弱

➢ 安全信任痛点

- 隐私安全风险高
- 用户信任建立难
- 生态规范构建慢

GUI 模型训练的痛点问题



真实场景性能差

- 迟延
- 精度

传统范式上限低

- 微调
- 强化

安全隐私无保障

- 数据
- 模型

评测体系不完善

- 中文
- 导航
- 动态

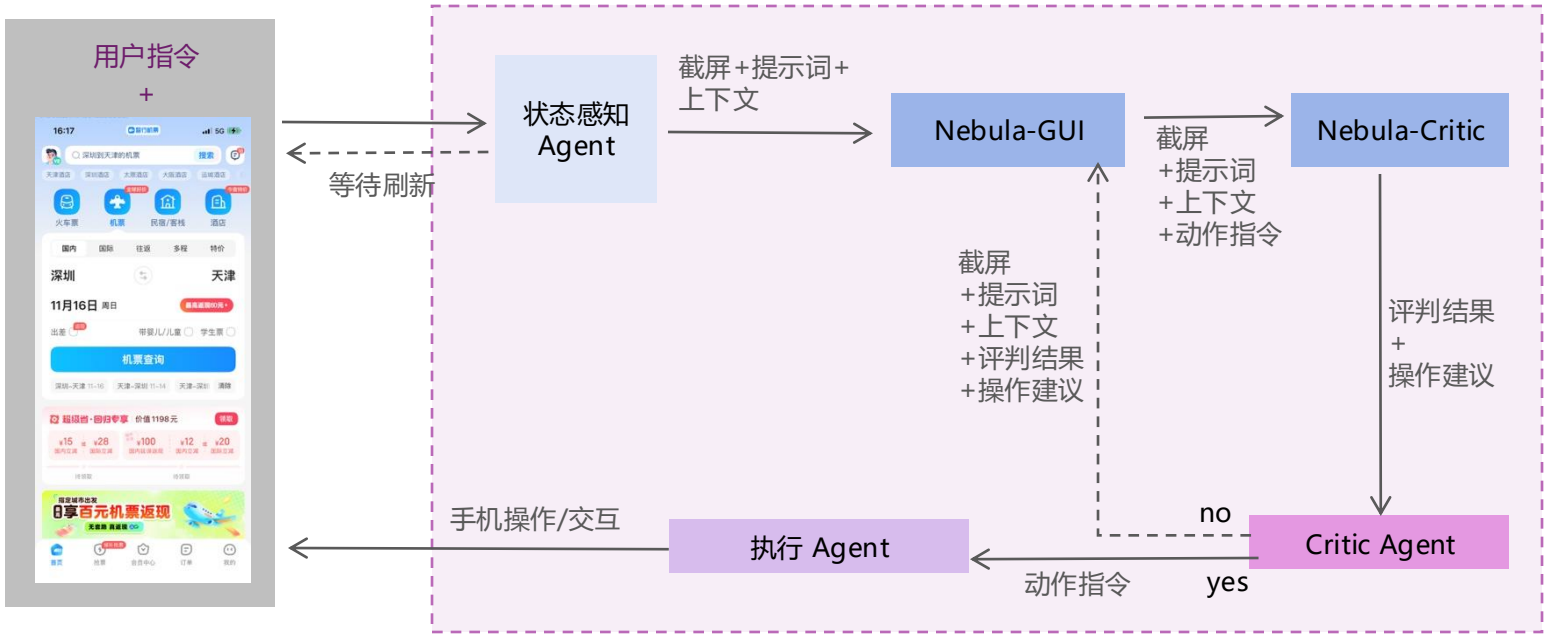
高质量数据缺失

- 量级
- 标注
- 轨迹

PART 03

整体方案

Nebula-GUI Agent 的整体方案



Nebula-GUI Agent

➢ 模块组成：

- Nebula-GUI 模型：负责用户指令的理解、拆分、给出规划路径和操作指令
- Nebula-Critic 模型：负责分析Nebula-GUI模型给出的决策是否正确
- 状态感知Agent：检测页面是否完成更新
- Critic Agent：分析评判结果
- 执行 Agent：将模型的输出映射为手机操作指令

Nebula-GUI 整体解决方案

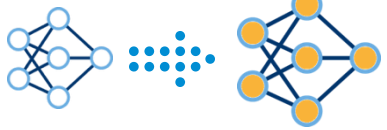
- 针对GUI Agent的痛点问题，我们从数据、训练、评测、推理打造全流程解决方案

数据

- 开发GUI数据标注工具
- 开发数据自动化生成系统
- 开发设备联邦管理系统
- 开发高效智能化数据pipeline

千万级高质量训练样本

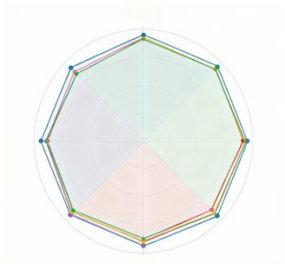
训练



- 使用7B VLM 模型
- SFT引入think with image
- GRPO强化引入定制化奖励
- 任务级在线强化学习

精度高、泛化强、速度快、可信赖

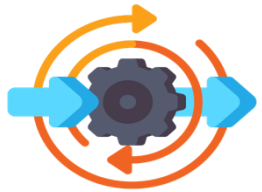
评测



- 构建多维评测集

指导模型精准、全面发展

推理



- 引入自适应思考模式
- 引入评判模型

快速、准确的推理能力

使用

精、准、快、稳

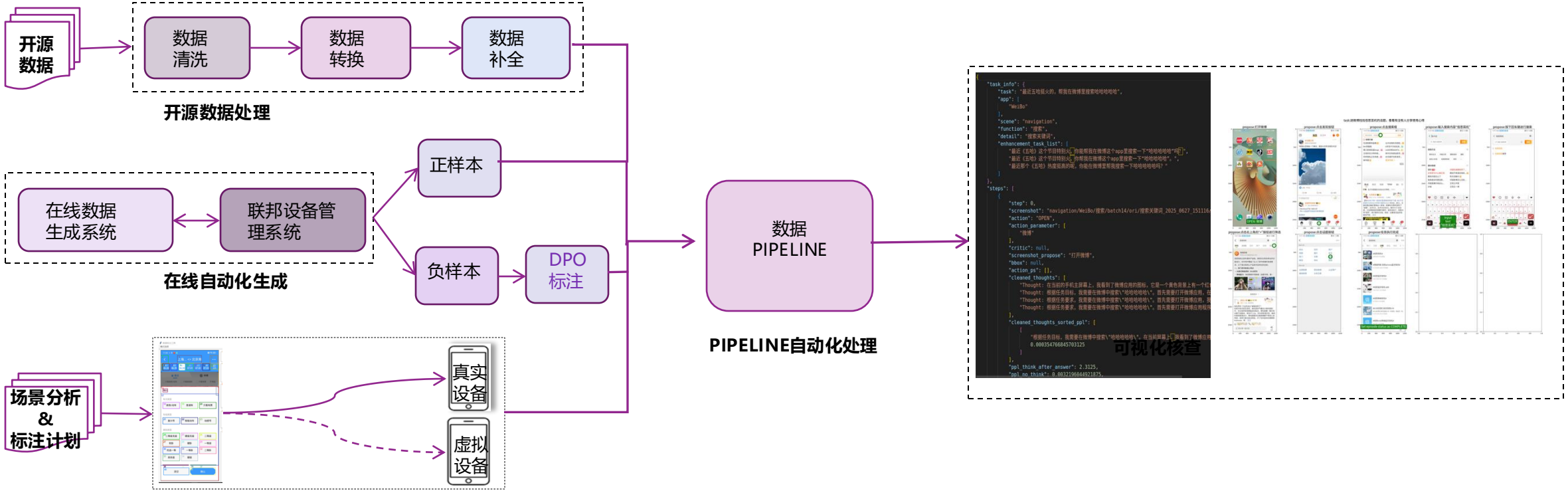
PART 04

具体实现

数据质量提升

核心技术创新点

- **一体化数采标注工具**：提升数采效率3倍以上，从机制上保证了标注的准确性
- **联邦设备管理系统**：物理和虚拟设备统一纳管，突破设备数量限制，为在线强化、数据飞轮提供了基础设施支撑
- **自动化轨迹生成系统**：在线自主探索，自动化生成训练样本，提供百万条覆盖50+ APP的导航数据
- **高效智能化数据Pipeline：**全链路自动化一键完成，每日可生成**万级**数据



数据质量提升--一体化数据采集标注工具

控件识别：自动识别所有UI控件进行画框、解决人工标注边框偏差、不全等问题

精准标注：自动获取UI控件的边界框和操作坐标，解决人工标注画框繁琐等问题

行为定义：支持标注多种操作类型，弥补开源action数据少的问题

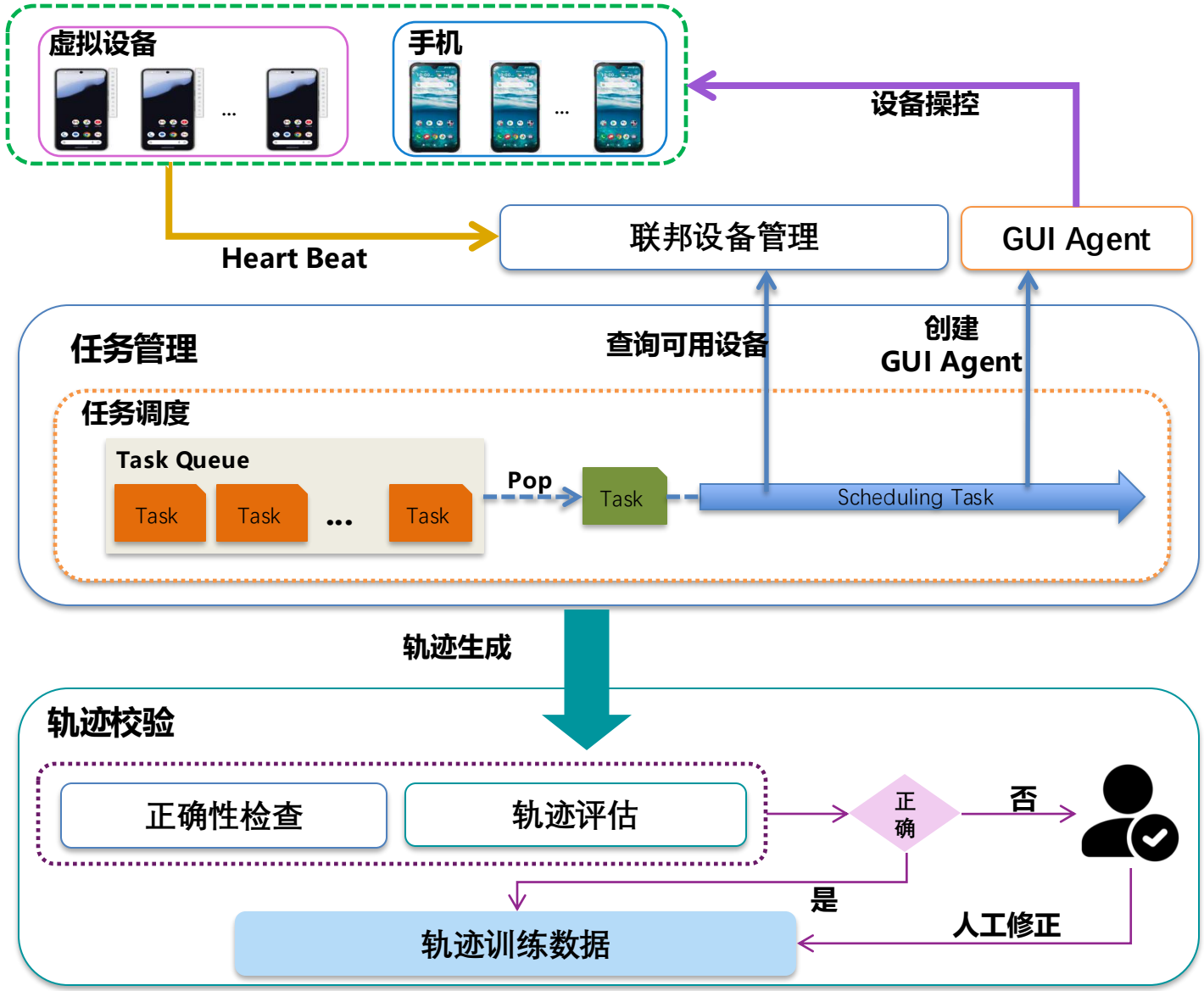
功能描述：支持控件功能描述标注，提升模型对于控件的功能理解

语义丰富：借助大模型产生更多的泛化指令，提供大量高质量单步数据

附加标注：支持标注思维链等信息，解决复杂轨迹路径模型学习难的问题

自动归档：自动汇总所有步骤信息（操作、截屏、任务信息）并生成指定数据格式。解决人工汇总流程繁琐、易出错的问题

数据质量提升--自动化轨迹生成系统



问题：屏幕数据标注对人力与设备资源有双重依赖

- 推出集任务调度、任务管理、联邦设备管理与GUI Agent于一体的数据自动化生成平台
- 大量虚拟设备解决了实体设备少的问题
- 实体设备解决了部分APP无法在虚拟设备操作的问题，两者互为补充
- 联邦设备管理有效解决了设备资源利用率低、管理分散的痛点。
- 大幅降低了对人工标注的依赖，人力成本减少3倍。

数据质量提升--高效智能化数据PIPELINE

➢ 核心功能

- 批量处理大量开源、自采数据，每天可生产**数万条**高质量数据，保证模型在更多APP上都有理解和操作能力
- 混合多个大模型和规则处理，能够得到准确率更高，任务泛化性更强的数据，提升模型性能

图片

规则筛选

文本清洗

标注

任务增强

思维链生成

思维链评估

指令泛化

Gemini

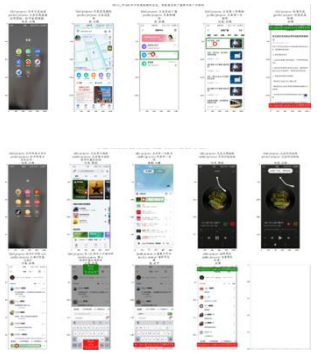
.....

✓ 为了找到用户的歌单，应该先选择用户个人页面;
✓ 要找到用户的歌单，得先从他的个人页面入手;
✓ 导航至用户个人页面，以访问其歌单;

轨迹流程可视化

指令模拟校验

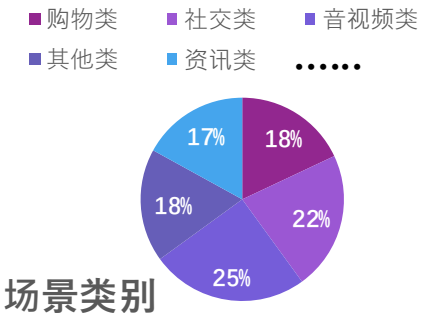
人工抽查审核



场景采样

步骤采样

多任务采样

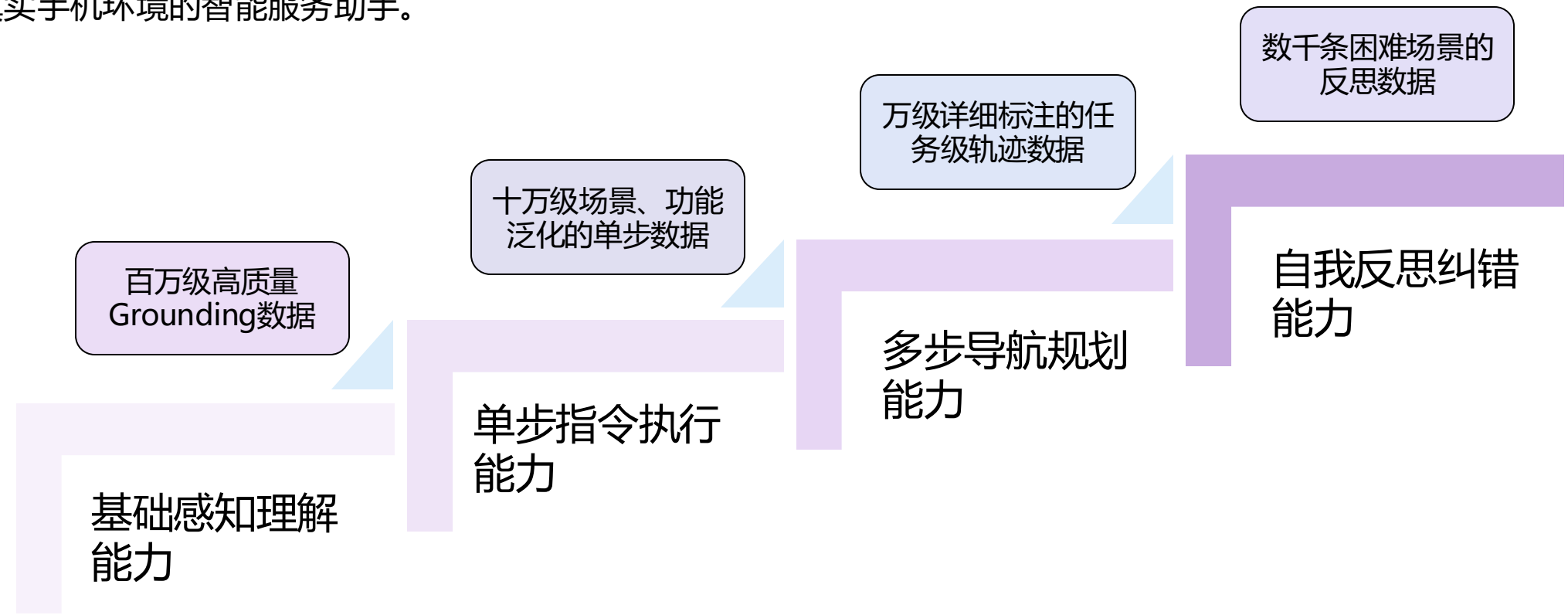


场景类别

数据清洗 → **数据增强** → **数据审查** → **数据采样**

训练方法改进--多阶段微调

- **基础操作鲁棒性显著增强：**模型对中文GUI页面，UI元素的动态变化具备了更强的理解能力，能够准确识别目标组件，有效抵御界面噪音干扰，大幅降低了单步操作的失败率。
- **长任务流程成功率有效提升：**规划能力的注入与自我纠错机制，使智能体能够像人类用户一样，在执行中监测状态，在偏离时回溯路径，保证了复杂多步任务的完成度。
- **从“实验室原型”迈向“商业可用”：**Nebula-GUI Agent从自由度较低、行为死板的demo级产物，进化为一个能够适应真实手机环境的智能服务助手。



训练方法改进--SFT引入think with image

优化一：基于图像Grounding思考

Q：点一杯少糖茉莉花香拿铁

文本内容和图像相关区域相互关联

think：“少糖”位于图中间位置(54, 2160, 523, 2198)，因此我需要点击(54, 2160, 523, 2198)

- 思维链中文本内容需要与图像相关区域进行关联，增强了一致性，解决了传统思考中以文本为中心的思维链易产生幻觉的问题；
- 优化后，验证集性能从88%提升至94%

优化二：基于图像操作思考

Q：手机的WIFI是否打开了？

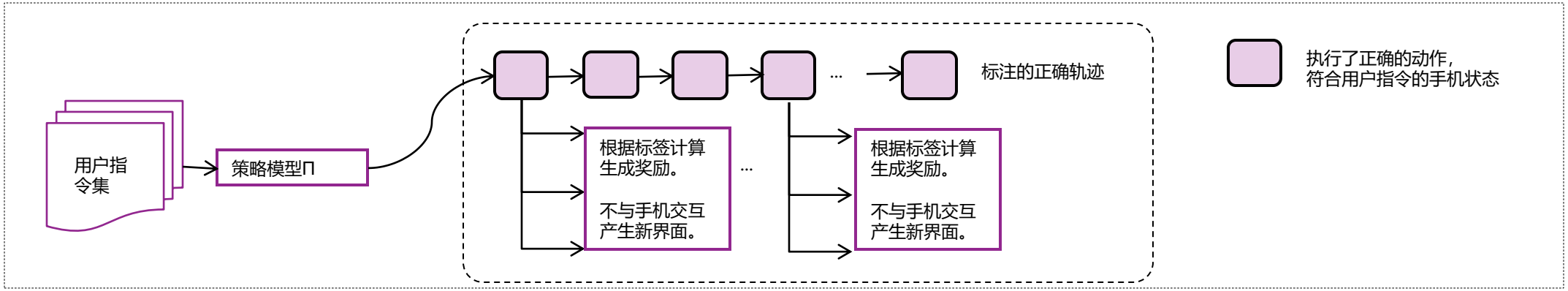
<TOOL_CALL>
ZOOM IN

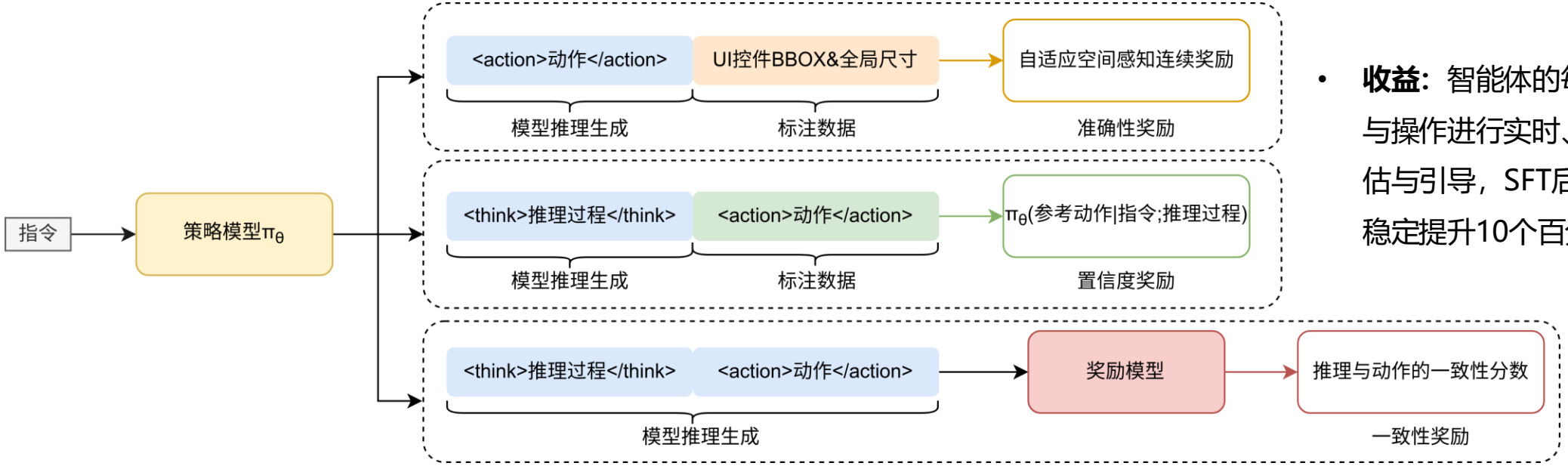
think：根据新的图片状态来看，当前的WiFi是已经链接的状态，说明用户的WiFi已经打开了，并且成功连上了WiFi。

- 通过对图像操作获取新的图像内容作为补充，解决了传统思考中图像一次编码导致细节理解不清楚的问题，进一步提升回复准确性
- 优化后，细节问题，模型能力提升8个百分点

训练方法改进--GRPO强化定制奖励

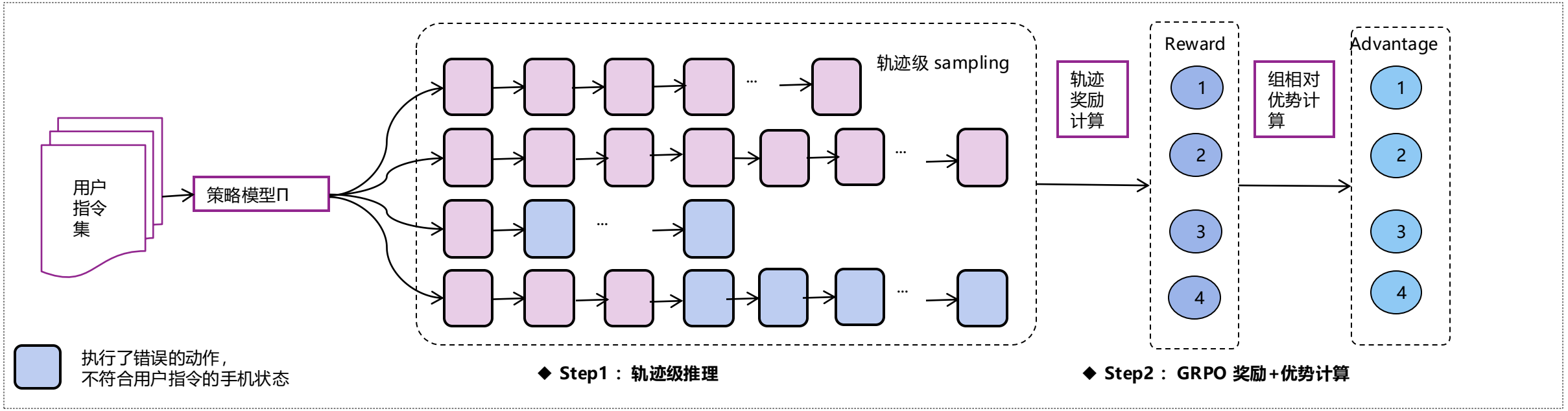
- 针对GUI agent在任务中奖励粒度粗糙的问题，我们改进了传统的GRPO离散奖励框架，设计了细粒度的连续性奖励信号

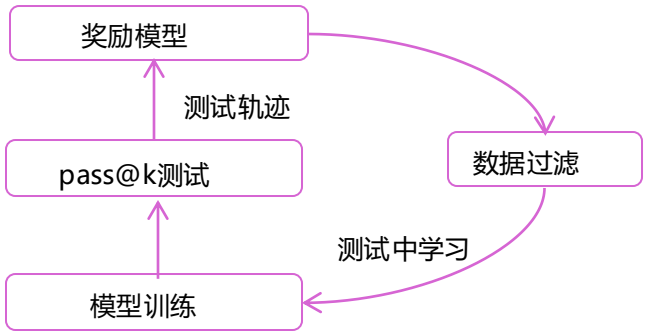




- **收益：**智能体的每一步推理与操作进行实时、精准的评估与引导，SFT后的指标能稳定提升10个百分点。

训练方法改进--任务级在线强化学习





任务级在线强化学习算法目标

改进点： 1. 引入在线RL范式，训练过程中能和手机环境进行真实交互

2. 引入轨迹级奖励，以任务是否成功作为依据

收益： 通过与环境持续交互，能够动态探索新状态，充分利用失败轨迹，从而提升模型对长尾场景的覆盖能力和泛化性能

精准评测--构建多维度评测benchmark

测评维度

- GUI基础感知能力
  - 页面理解
  - UI 功能理解
  - UI 状态理解
  - UI Grounding
    - ScreenQA-Short
    - ScreenQA-Complex
    - ScreenSpot-v2
    - CAGUI-Grounding
    - 基于主流APP制作的基础感知类测试集
- GUI控件单步操作能力
  - 操作指令执行能力
  - 指令与元素关联能力
    - AndriodControl
    - AMEX
    - 基于主流APP人工标注数据制作的离线单步指令测试集
- 多步任务执行能力
  - 简单任务执行能力
  - 复杂任务执行能力
    - CAGUI-Task
    - AndriodWorld
    - AndriodLab
    - 主流 APP 离线任务测试集
    - 主流 APP 在线测试集，人工审核

测评讲解demo页面

离线+在线，感知+操作+导航，多个维度测试，保证模型在真实场景的高精度

推理速度优化--引入自适应思考模式

任务：帮我在12306上购买一张明天7点左右去南京的高铁票

STEP 1

简单操作
无思考

OPERATION: 打开12306 APP程序

ACTION: OPEN: ["铁路12306"]

任务：帮我在12306上购买一张明天7点左右去南京的高铁票

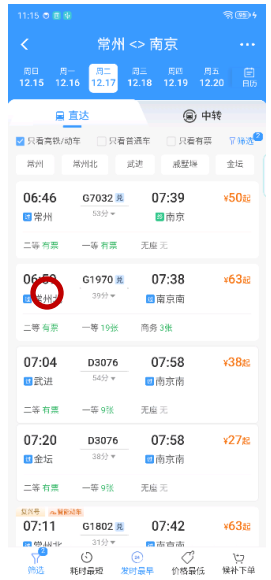
STEP 10

复杂操作
有思考

THINK: 根据任务要求，我需要为购买9月6日7点从天津到宁波的二等F座车票。在当前页面，我看到了多个列车车次，其中6点59分的G1970车次正好满足要求，我应该点击该列车次进行购买

OPERATION: 点击G1970车次的始发站

ACTION: CLICK: [216, 483]



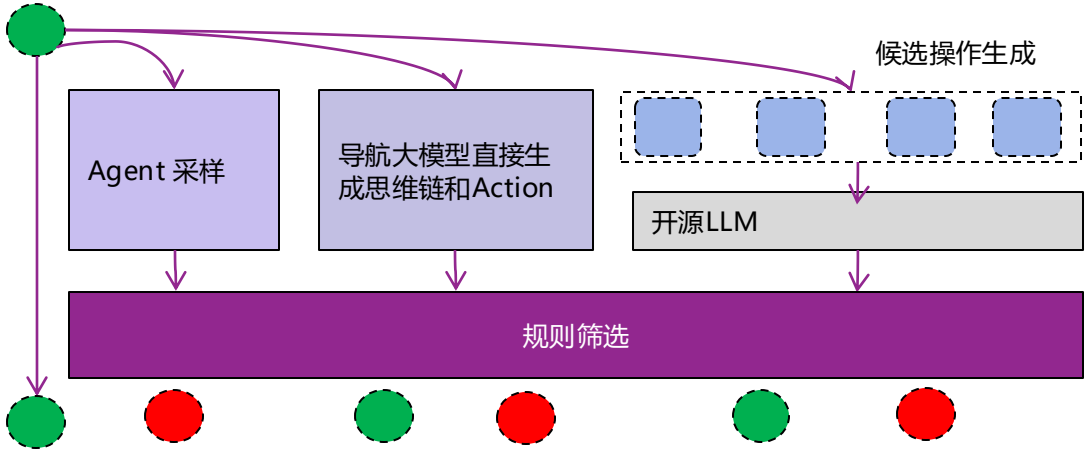
SFT: no think+think (all task) → 测试：pass@k → 强化：
简单任务：no_think 格式输出
复杂任务：think 格式输出 → 自适应思考模式

兼顾模型输出质量和执行效率：简单步骤直接输出操作结果，复杂步骤输出思考过程，最终推理时延在 1s/步

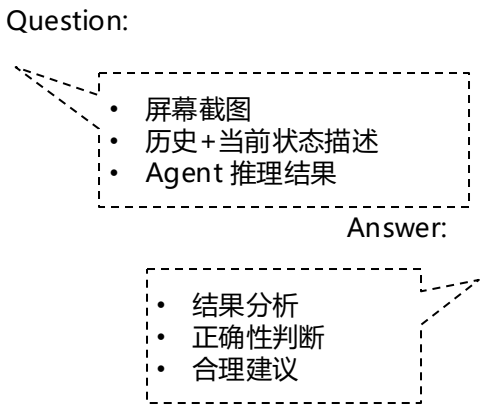
推理质量保障--引入评判模型

口 评判模型(Critic Model)

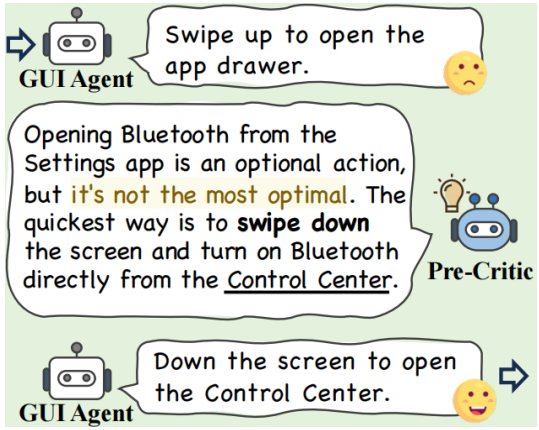
- 在执行的过程中，为Agent模型提供动作执行预校检，包括**正确性判断、结果分析、合理建议**，提升系统准确率和对Agent模型的容错能力。
- 策略模型综合考虑历史、当前屏幕状态以及批评模型建议，输出下一步动动作。
- 优化prompt设计，避免评判模型主导操作流程。



◆ **Step1： 数据采集（正负样本构建）**



◆ **Step2： 模型训练**



◆ **Step3： 在线部署**

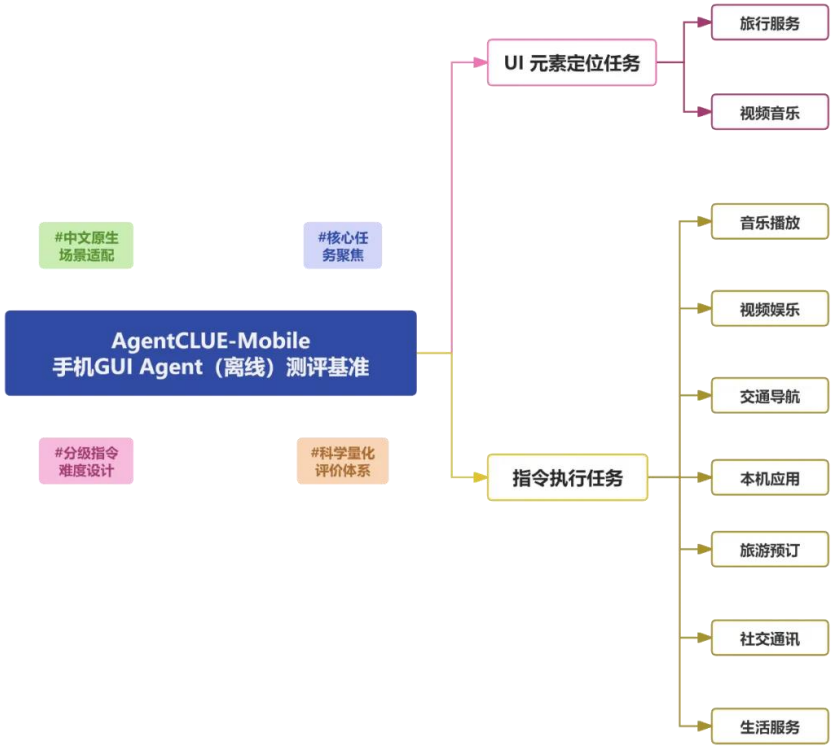
性能：14个主流App TOP5场景成功率95%

APP	模型能力	Nebula GUI agent	开源方案 1	开源方案 2	开源方案 3
拼多多	购买商品、浏览商品、收藏商品、分享商品给微信好友、写好评	100%	-（不支持）	66.7%	28%
携程	购买旅游景点门票、线路规划、旅游攻略查询、订票、订单评价	94.7%	12.5%	66.7%	-（不支持）
抖音	搜索视频、播放指定模块、团购、写好评；视频的点赞、收藏等简单操作、查看指定模块、抖音商城	94%	60%	53.8%	33%
大众点评	指定内容搜索、团购、写好评、查看关键信息、筛选景点	92%	49%	60%	17%
微博	指定内容搜索、点赞、收藏、转发、评论、分享、用户相关操作	100%	53%	66.7%	40%
小红书	指定内容搜索、点赞、收藏、评论、分享、查看、用户操作	96%	72%	100%	45%
爱奇艺	播放指定内容的视频、浏览影片、播放设置	100%	100%	-（不支持）	4%
京东	购买商品、浏览商品、管理购物车、写好评、点外卖、商品分享	93%	73%	22.2%	8%
网易云音乐	播放音乐、收藏音乐及歌单、播放设置、歌曲分享、收藏	100%	-（不支持）	100%	11%
高德	出行导航、查看附近、出行查询、导航设置、打车	90.68%	81.25%	85.7%	-（不支持）
淘宝	购买商品、浏览商品、点外卖、写好评、购物车管理、商品分享	93.75%	62.5%	66.7%	23%
微信	发朋友圈、发消息、发红包、搜索公众号或服务号、创建群聊	95%	-（不支持）	100%	45%
美团	外卖、写好评	94%	87%	37.5%	40%
12306	订票、查询车票	90%	-（不支持）	50%	33%
平均准确率		95%	46%	62.5%	23.3%

单步推理迟延：Nebula-GUI Agent 1s，开源方案1 4.3s，开源方案2 5s，开源方案3 3s

性能：AgentCLUE-Mobile榜单银牌

- **Agentclue-Mobile 第三方打榜:**
  - 我司推出的Nebula-GUI Agent，以7B 策略模型，在第三方权威评测AgentCLUE-Mobile中荣获银牌
- **商用落地**
  - 一句话订餐、一句话写好评（包含大众点评、美团、京东等TOP6应用内），12306订票已在我司Z70 Ultra、Z80U和红魔手机落地商用。



AgentCLUE-Mobile
手机 GUI Agent（离线）基准测评总榜概览

排名	模型	机构	总分	UI元素定位得分	指令执行得分	使用方式
🥇	GLM-4.5v	智谱AI	**90.75**	**93.60**	**90.04**	API
🥈	MiMo-VL-7B-RL-2508	小米集团	**90.01**	85.20	**89.22**	模型
🥉	Doubao-Seed-1.6-thinking-250715	字节跳动	**84.59**	**94.78**	82.04	API
4	Nebula-GUI	中兴通讯	**84.38**	93.17	**82.19**	API
5	AgentCPM-GUI	面壁智能	**79.87**	76.40	80.74	模型
-	Gemini-2.5-pro	Google	**71.86**	59.20	75.03	API
4	GUI-Owl-7B	阿里巴巴	**69.62**	**96.80**	62.82	模型
5	qwen2.5-vl-7b-instruct	阿里巴巴	**66.84**	92.00	60.55	模型
6	qwen2.5-vl-3b-instruct	阿里巴巴	**61.53**	90.70	54.24	模型
7	MiniCPM-V4.5-8B	面壁智能	**56.85**	84.00	50.07	模型
8	ui-tars-1.5-7b	字节跳动	**46.40**	79.60	38.10	API
-	Gemma3-4B-it	Google	**27.94**	5.40	33.58	API
9	DeepSeek-VL2-tiny	深度求索	**23.32**	6.00	27.65	模型

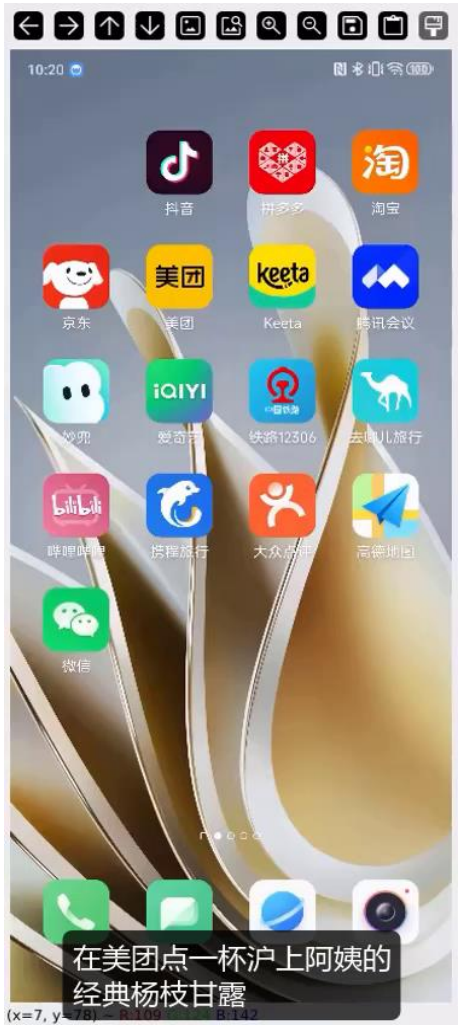
排名计算方式说明：为减少波动影响，榜单将分差1分值内的模型视为并列排名。国外模型及补测模型的旧版本不参与排名，只做参考。
数据来源：SuperCLUE，2025年10月17日。

AgentCLUE-Mobile
手机 GUI Agent（离线）基准测评应用场景榜单

排名	模型	机构	指令执行得分	视频娱乐	音乐播放	交通导航	本机应用	办公资讯	旅行预订	生活服务
🥇	GLM-4.5v	智谱AI	**90.04**	**92.00**	**86.00**	**96.00**	**86.00**	**86.00**	**93.00**	**91.00**
🥈	MiMo-VL-7B-RL-2508	小米集团	**89.22**	**90.00**	**88.00**	**92.00**	**87.00**	**85.00**	**90.00**	**92.00**
🥉	Nebula-GUI	中兴通讯	**82.19**	**86.00**	76.00	74.00	**83.00**	**87.00**	**86.00**	83.00
4	Doubao-Seed-1.6-thinking-250715	字节跳动	**82.04**	85.00	**79.00**	80.00	82.00	80.00	**86.00**	**84.00**
5	AgentCPM-GUI	面壁智能	**80.74**	82.00	78.00	**85.00**	73.00	**85.00**	84.00	78.00
-	Gemini-2.5-pro	Google	**75.03**	74.00	71.00	68.00	74.00	75.00	84.00	81.00
4	GUI-Owl-7B	阿里巴巴	**62.82**	62.00	62.00	63.00	60.00	63.00	65.00	64.00
5	qwen2.5-vl-7b-instruct	阿里巴巴	**60.55**	62.00	61.00	61.00	54.00	60.00	65.00	60.00
6	qwen2.5-vl-3b-instruct	阿里巴巴	**54.24**	57.00	56.00	58.00	37.00	54.00	60.00	55.00
7	MiniCPM-V4.5-8B	面壁智能	**50.07**	49.00	42.00	49.00	46.00	45.00	45.00	66.00
8	ui-tars-1.5-7b	字节跳动	**38.10**	35.00	39.00	33.00	38.00	38.00	40.00	43.00
-	Gemma3-4B-it	Google	**33.58**	32.00	35.00	35.00	32.00	33.00	35.00	34.00
9	DeepSeek-VL2-tiny	深度求索	**27.65**	31.00	31.00	29.00	24.00	30.00	21.00	25.00

排名计算方式说明：为减少波动影响，榜单将分差1分值内的模型视为并列排名。国外模型及补测模型的旧版本不参与排名，只做参考。
数据来源：SuperCLUE，2025年10月17日。

性能：实例展示



美团点外卖

点评收藏、打卡+好评

订机票

订酒店

PART 05

总结与展望

总结与展望

- 我们通过高效率的数据工程、高质量的监督微调、双层强化学习、全方位的能力评估和模型推理优化等一系列改进措施，目前Nebula-GUI Agent能够在手机大多数场景上表现良好，但落地商用方面还有以下问题亟需解决：

  1. 用户信任：构建用户信任与可控性的基石
  2. 场景智能：实现场景化分析与动态决策的能力
  3. 个性适应：满足用户偏好与习惯的操作

- 未来中兴通讯将在GUI Agent的研发上不断突破，全力推动该技术在智能办公、软件开发、自动化流程等领域的深度渗透与规模化应用，让尖端技术真正赋能各行各业，创造可衡量的巨大价值

科技生态圈峰会 + 深度研习

——1000 + 技术团队的共同选择

KEYLINK ing

AiDD 峰会

AiDD峰会 上海站
AI+研发数字峰会
时间：2026.05.22-23

AiDD峰会 北京站
AI+研发数字峰会
时间：2026.08.21-22

AiDD峰会 深圳站
AI+研发数字峰会
时间：2026.11.20-21

AiDD峰会详情

AiDD × AI+产品峰会

AiDD × 上海站
AI+产品创新峰会
时间：2026.01.16-17

AiDD × 北京站
AI+产品创新峰会
时间：2026.08.23-24

产品峰会详情

AiDD × K

2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会

Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞家与需求发现

论坛2: AI原生产品战路与商业模式重构

论坛3: AgenticAI产品创新与交互设计

Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

2-hour Speech: 回归本质

用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，仍然考验“看见”的“同理心”

Panel 1: 出海前瞻

“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式

Panel 2: 失败复盘

为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长

Reinventing Products with Intelligence, Driven by Data

80+ 业界大咖 汇聚产品大咖的真知灼见

40+ 创新案例 开拓全新AI+产品视角

10+ 分论坛 涵盖产品到商业增长的全链路

800+ 听众规模 共同探索产品的智能重构

扫码查看会议详情

第8届AI+研发数字峰会
AI+ Development Digital Summit

感谢聆听！

扫码领取会议PPT资料