计算机使用代理

2025年2月11日 1 分钟阅读

#OpenAI #Operator #论文 #AI

> *计算机使用代理*

各种操作电脑的AI代理

ChatGLM AutoGLM

OpenAI Operator

Byte Dancing OmniHuman-1

Anthropic Claude

Meta Agent

OpenAI Operator系统概述

一、引言

Operator是OpenAI的计算机使用代理（CUA）模型研究预览版，融合GPT-4o视觉与强化学习推理能力，可像人类一样与图形用户界面交互，能执行如网购、预订等日常任务，但也带来如提示注入攻击等风险。OpenAI为此构建多层安全体系，本文详述其测试与部署策略、风险识别与缓解措施。

二、模型数据与训练

采用监督学习与强化学习结合方式，基于多种数据集训练，包括公开数据和人工创建数据，使模型掌握计算机屏幕感知及输入控制技能，实现如推理、纠错和适应意外情况等高级能力。

文章信息

字数

阅读时间

发布时间

更新时间

标签

#OpenAI #Operator #论文

三、风险识别

（一）政策制定

依任务和行动风险程度分类，对如金融交易等高风险操作设保障措施，如关键步骤人工监督、行动前确认，限制股票交易等任务，确保用户对模型操作可控。

（二）红队测试

先内部后外部红队测试。外部红队来自二十国、懂二十多种语言，测试模型能力、安全措施及对抗输入能力，在模拟环境查找漏洞，为强化安全保障提供依据，当前Operator以研究预览版向小范围用户部署以便监测。

（三）前沿风险评估

依OpenAI准备框架评估，Operator在说服和网络安全类别继承GPT-4o风险等级（中、低）。在生物风险工具（CBRN）和模型自主性方面，因视觉输入和光标输出限制，预缓解模型风险均为低，如生物风险工具任务成功率仅1%，模型自主性主要任务得分不超10%。

四、风险缓解

（一）有害任务

用户受OpenAI使用政策约束，模型在训练中拒绝有害任务，内部评估对新代理危害任务拒绝率达97%，超GPT-4o。系统限制访问有害网站，部署后将自动和人工审查监测滥用，跟踪缓解效果并优化。

（二）模型错误

未缓解模型在测试中有13处错误，部分可逆。通过确认、主动拒绝、观察模式等措施降低风险，如确认操作使风险降低约90%，特定评估集中主动拒绝高风险任务召回率94%，在敏感网站观察模式可暂停执行待用户监督。

（三）提示注入

相比未缓解模型的62%和仅提示模型的47%，缓解后模型对31种提示注入场景敏感度降至23%，且有提示注入监测器，对红队测试77次尝试召回率99%、精度90%，可快速更新应对新攻击，其他风险缓解措施也对其有防范作用。

五、局限与未来工作

虽采取措施，但因现实复杂和威胁动态仍有局限。未来计划包括收集反馈提升模型质量、逐步扩大用户范围、开放API并确保安全、持续提升安全性和策略合规性，各团队将持续协作改进。

OpenAI Operator OpenAI Operator system card

Byte Dancing midscenejs

用于通过人工智能以自然语言自动操作网页的开源软件开发工具包。Midscene.js将用户界面自动化转变为一种愉悦的体验。通过引入人工智能功能，Midscene使用户能够使用自然语言与网页无缝交互、查询和断言。

midscenejs

ChatGLM(智谱) AutoGLM

AutoGLM

AutoGLM 论文

智谱的 **AutoGLM** 是一款非常有突破性的自主智能体，它在实现跨平台任务方面主要依赖于以下几个关键技术和设计理念：

1. 基于图形化用户交互界面（GUI）的操作

AutoGLM 模拟人类操作：与传统的 API 调用不同，AutoGLM 通过模拟人类的屏幕操作实现跨平台任务。例如，它可以像人类一样点击按钮、填写表单、滑动屏幕等。

优势：

避免了 API 标准化不足的问题：由于许多网站和 APP 的 API 缺乏统一标准，传统 API 集成容易因应用更新而失效。而 GUI 模拟只要界面保持用户可理解，就能持续工作。

灵活适配：无需依赖开发者提供专门的 API，AutoGLM 可以适配更多应用场景。

2. 多模态模型支持

视觉多模态支持：AutoGLM 基于智谱的多模态模型（如 GLM-PC 和 CogAgent），能够理解和分析屏幕上的视觉信息。这意味着它不仅能读取文本，还能识别图像、按钮、图标等内容。

任务规划与执行：通过视觉分析和逻辑推理，AutoGLM 可以规划任务步骤，并将任务分解为多个具体的操作。

3. 任务链的自动调度与执行

任务链处理能力：AutoGLM 能够自主完成超过 50 步的复杂任务。例如，在购买火锅食材的场景中，它可以从选择商品到支付全程无人干预完成。

跨应用协作：用户只需发出简单指令（如"买咖啡"），AutoGLM 就能在多个应用之间自动切换，完成从选择商品到支付的整个流程。

4. 智能化的个性化决策

"随便模式"：AutoGLM 能够分析用户的偏好和历史行为，主动为用户做出决策。例如，当用户犹豫时，它可以根据用户的习惯推荐最优选项。

实时学习与优化：通过不断学习用户的操作习惯，AutoGLM 提升了任务执行的准确性和效率。

5. 核心技术支持

高效的多平台适配：AutoGLM 支持手机、浏览器、PC 等多种设备，具备跨平台操作能力。

端云一体架构：通过与芯片厂商（如高通、英特尔）的合作，AutoGLM 优化了端侧性能，使其适配多种设备，并通过云端计算增强复杂任务的执行能力。

示例场景：跨平台任务执行

用户指令：
"帮我订今晚的火锅外卖。"
AutoGLM 执行步骤：

打开外卖 APP，搜索火锅店铺。

根据用户历史订单推荐火锅套餐。

切换到优惠券 APP，自动领取可用优惠券。

返回外卖 APP，应用优惠券并完成支付。

在社交媒体上分享订单（可选）。

整个过程无需用户干预，体现了 AutoGLM 的强大跨平台能力。

总结

AutoGLM 的跨平台任务能力主要依赖于 **模拟人类操作的 GUI 技术**、**多模态模型的支持**、**任务链自动调度** 和 **个性化决策**。这些技术的结合使其能够灵活适配多种应用场景，无需依赖开发者提供专用 API，从而实现真正的跨平台任务执行。