

图形用户界面智能体的执行、验证与风险：平台感知综述

武宇¹，杨易²

¹武汉大学人工智能学院，中国武汉市，430072

²浙江大学人工智能省部共建协同创新中心ReLER实验室，中国杭州市，310027

摘要：图形用户界面（GUI）智能体已广泛应用于通用数字自动化。大语言模型、视觉语言模型和多模态基础模型现在能够解释屏幕截图，遵循自然语言指令，并在网站、移动应用和桌面操作系统中执行具备定位依据的操作。然而，尽管这些场景具有不同的执行条件，现有研究仍常常将其归入同一标签之下。本文通过一个共同的技术栈——包括观察、定位、规划、记忆、执行和验证——综述了GUI智能体的代表性框架、模型、数据集和测试基准。我们主张将Web智能体、移动端智能体和桌面/操作系统智能体视为不同的操作体制，因为它们在可观测性、动作语义、隐藏状态依赖、执行成本和副作用风险方面存在差异。描述了该领域从下一步动作预测向可靠执行的转变，其中结果验证、执行效率和可信性受到越来越多关注。最后，探讨了面向可靠计算机使用的未来研究方向，包括信念状态跟踪、自适应感知、GUI与工具混合执行以及人类监督。

关键词：图形用户界面智能体；操作体制；可信评测；后置条件验证
<https://doi.org/10.1631/ENG.ITEE.2026.0103>