我们常说AI agent越来越聪明了，但聪明到底怎么量化？如何判断一个AI agent到底好不好用，其实这就像面试一位员工，你要看他能不能干活，干得好不好？遇事稳不稳？是不是靠谱？今天我们用一个八字口诀：诚忧快稳、自推能守，来系统拆解评估AI agent的八个核心指标。

**1、诚（任务完成度task success rate）：**

这是最基础也最重要的标准，agent有没有完成他该做的任务，比如你让他查找一份财报生成一段代码，他有没有成功完成，我们会统计他在测试集上的任务完成比例来衡量。

**2、忧（结果质量quality of outcome）：**

任务完成了，那完成得好不好？比如生成的代码能不能跑？写的总结有没有条理？这通常通过人工打分，自动化指标如代码通过率，文本的ROUGE分数来判断。

**3、快（效率efficiency）：**

一个agent慢吞吞，操作复杂，就算再聪明也不实，用我们会评估他完成任务所花的时间、步骤数量，调用工具或语言模型的频率，以及整体token消耗。

**4、稳（鲁棒性robustness）：**

一个真正可靠的agent必须能扛住各种干扰，例如我们会故意给他一些模糊甚至错误的信息，看他会不会崩溃，或者故意让外部工具故障，测试他是否能优雅应对。

**5、自（自主性autonomy）：**

优秀的agent应该像一个能独当一面的员工，不需要我们实时介入，评估时我们会观察他解决问题时是否频繁需要人工协助。

**6、推（推理能力reasoning ability）：**

这指的是agent是否能理解任务、拆解问题、并规划多步操作，例如在react框架中，我们会特别分析他的thought思考过程，看他是否具备逻辑性和条理性。

**7、能（泛化能力generalization）：**

真正强的agent不只是擅长他见过的任务，更能应对新问题新环境，我们会在他没见过的数据或情境中测试他的应变能力。

**8、守（安全与对齐safetyalignment）：**

AI再强也必须守规矩，我们会设计一些挑战性的安全场景，测试他是否遵循伦理原则，是否可能产生有害或偏离预期的行为。

目前，像agentbench toolbench webarena GAIA等评估框架，正被广泛应用于上述指标的综合测试，它们帮助我们更系统更全面地了解agent的真实能力，而不仅仅是表面看起来很智能。

总结一下一个优秀的AI agent不只是能干活，还要干得好，干得快，干得稳，更要会思考，能适应，守规矩。