精选快讯　软件快讯　2026年11月18日 05:49

Anthropic警告AI安全测试过于简单，建议增加多轮对话评估

要点

- Anthropic在周五发布的博客文章中警告称，当前的AI安全测试过于简单，难以有效评估模型在复杂场景下的表现。
- 数据显示，Anthropic API上单轮评估的无害行为响应率一致保持在94%至98%之间，而多轮响应则显示出更宽的范围。
- Anthropic强调，通过增加多轮对话评估是提高覆盖率的关键措施。

Anthropic在周五发布的博客文章中警告称，当前的AI安全测试过于简单，难以有效评估模型在复杂场景下的表现。该公司指出，虽然模型擅长立即识别并拒绝直接的有害请求，但在面对多轮对话中逐渐升级的隐蔽有害意图时，其应对能力存在明显短板。Anthropic认为，由于模型能力已超出当前安全评估的测试范围，现有的单一轮次任务虽然易于构建，但缺乏现实性，无法完全展示模型能力的广度。

数据显示，Anthropic API上单轮评估的无害行为响应率一致保持在94%至98%之间，而多轮响应则显示出更宽的范围。为了解决这一问题，Anthropic建议调整任务难度，并引入覆盖现实测试用例的领域专家。同时，公司提议建立具备采样、评分和分析模型响应技能的专门测试团队，以更全面地展示模型能力并减少测试饱和。

Anthropic强调，通过增加多轮对话评估是提高覆盖率的关键措施。这种评估方式能够更真实地模拟用户与AI交互的过程，从而揭示模型在长期对话中可能出现的潜在风险。公司认为，仅依赖简单的单轮测试已不足以反映现代AI系统的实际安全水平，必须通过更复杂的评估框架来确保模型在面对隐蔽且渐进的有害意图时仍能保持稳健。

报道来源：bankinfosecurity、govinfosecurity

来源网站：https://www.inews.today/