澳洲幸运十官方开奖结果51

EMCm7DuGMf9lBRLV

澳洲幸运十官方开奖结果51AI 助手 Claude 的“内心世界”：Anthropic 新研究解密其价值观

IT之家 4 月 22 日消息，Anthropic 公司于周一发布了一项名为“Values in the Wild”的研究， 深入剖析了 AI 助手 Claude 在实际用户交互中的价值观表达。

研究团队从 Claude.ai 的 Free 和 Pro 用户中，收集了 2025 年 2 月 18 日至 25 日的 70 万条匿名对话数据，主要涉及 Claude 3.5 Sonnet 模型。团队经过筛选，聚焦于需要主观解读的对话，最终保留了 308210 条交互进行深入分析。

研究采用隐私保护框架 CLIO，确保数据处理中剔除私人信息，并设置了严格的聚合标准（如每组数据需包含超 1000 名用户），以保护用户隐私。

IT之家援引博文介绍，在 CLIO 框架下，Anthropic 利用自有语言模型提取了 Claude 表达的价值观，共识别出 3307 种 AI 价值观和 2483 种人类价值观。经人工验证，AI 提取的价值观与人类判断高度一致（一致率达 98.8%）。

这些价值观被归类为五个主要类别：Practical（实用性）、Epistemic（知识性）、Social（社会性）、Protective（保护性）和 Personal（个人性）。

其中，实用性和知识性价值观占主导，超过半数案例体现效率、质量或逻辑一致性等特质。

研究还发现，Claude 的价值观与 Anthropic 的 HHH 设计目标紧密相关，例如“用户赋能”对应 Helpful，“知识谦逊”对应 Honest，“患者福祉”对应 Harmless。

报告中还检测到“支配性”和“无道德性”等少量负面价值观，可能与用户尝试“越狱”模型有关。

研究揭示，Claude 的价值观表达并非一成不变，而是高度依赖具体情境。例如，在提供关系建议时，Claude 强调“健康界限”；讨论历史事件时，则注重“历史准确性”。

此外，Claude 在回应用户明确表达的价值观时，通常采取支持态度，在 43% 的相关交互中强化用户框架，甚至“镜像”用户价值观（如“真实性”）。

相比之下，Claude 较少“重塑”用户价值观（占比 6.6%），多见于个人福祉或人际关系讨论；而直接抵制用户价值观的情况更少（5.4%），通常发生在用户请求不道德内容或违反使用政策时。

奥洲10开奖号码

168澳洲5开奖结果历史

押单双最聪明三个公式

澳洲幸运10号码开奖结果是什么

168澳洲幸运10开奖计划

众赢国际版计划软件官网

买13458和02679技巧

澳洲10开奖网

168澳洲幸运10官网历史查询

168飞艇计划全天预测永久免费

澳洲10开奖号码查询

极速快三彩票一秒一开

澳洲幸运10真实吗

澳洲幸运10怎么玩胜算大

精准计划导师带赚包赔

7码2期雪球技巧心得

新澳门彩网站APP

168飞艇开奖官网开奖记录

澳洲幸运10官方开奖结果