

Leaderboard Report

xbench-DeepResearch

时间：2025 年 5 月

自主规划(Planning) -> 信息收集(Search) -> 推理分析(Reasoning) -> 总结归纳(Summarization) 的深度搜索能力，是 AI Agent 通向 AGI 的核心能力之一。然而，这一能力的复杂性也为评估工作带来了更高的挑战。当前，业界主流评测集侧重于基座模型的能力评估，高质量的 Agent 评测集相对稀缺。为了更好的考察 Agent 的深度搜索能力，红杉中国推出并开源了 xbench-DeepSearch 评测集，具备以下特点：

- **针对 Agent 设计：**题库中所有题目都需要 Agent 综合运用规划+搜索+推理+总结的端到端能力来解决。现有的知识搜索类基准测试（如 SimpleQA）主要测量模型检索简单事实的能力，不依赖高阶的规划+推理能力，对于当下模型来说过于简单，评测分数早已饱和。
- **专注深度搜索能力：**与 GAIA 等综合评测集不同，xbench-DeepSearch 定位深度搜索能力的评估，在题库设计时特别针对搜索空间的广度和推理深度进行了充分考量，帮助 Agent 开发者更精准地拆解 Agent 能力维度，快速定位性能瓶颈和优化方向。
- **适配中文互联网环境：**由于搜索与本地内容的信源质量高度相关，相比于同样定位深度搜索能力的 BrowseComp 评测集，xbench-DeepSearch 弥补了其中文语境搜索题库不足的弱点。
- **全新出题人工验证：**所有题目经由来自各行各业的专家人工出题，并由博士生交叉验证，力求题目的新颖性和主题的多样性，答案的正确性和唯一性，方便自动化评测。
- **持续更新长期维护：**每月榜单中持续汇报最新模型的能力表现，每季度至少更新一次评估集。同时，为了避免刷榜行为影响评测的公正性，我们在内部维护了一个闭源的黑盒版本，如果开源和闭源的排名相差较大，我们将会从榜单中移除相关排名和分数，以保证榜单结果的可信度。

如果您是 Agent 产品开发者，希望在榜单中加入您的优秀产品，或希望使用最新版本的 xbench 评测集来验证您的产品效果收集反馈，欢迎联系我们并提交产品的公开版本访问链接，我们会在约定的时间内完成评测任务，并将结果及时反馈给您。

xbench-DeepSearch Top 5 Distinguished Leaderboard

	模型/Agent 名称	UI 界面上 所选模式	公司	准确率 区间	评测时间	评测网页链接
1	o3	Search	OpenAI	65+	2025.05	https://chatgpt.com/
2	o4-mini-high	Search	OpenAI	60+	2025.05	https://chatgpt.com/
3	Doubao	Deep Think	ByteDance	50+	2025.05	https://doubao.com
	Grok-3	DeeperSearch	xAI		2025.05	https://grok.com
	Gemini 2.5 Pro (preview)	Deep Research	Google		2025.05	https://gemini.google.com

2025-05 xbench-DeepSearch 榜单说明：

- 我们观察到，众多开发者已着手运用模型的 AI Search 能力开发各类应用产品。其中，部分为我们已知的创业产品，部分尚处于信息收集阶段。为保证本次榜单的完整性与参考价值，本次测评与分数汇报仅面向大模型公司。
- 我们即将开源本测试集的白盒版本，以供广大开发者和创业者使用，并自行汇报分数。后续，我们将陆续收录白盒汇报中的优秀模型，开展黑盒验证工作，并在后续的黑盒 Leaderboard 中进行第三方汇报。
- 所有产品均开启联网功能进行评测。由于大部分 Agent 产品都有多种模式可以选择，表中标明了评测时 UI 界面上所选择的模式。
- 考虑到大模型的回复存在随机性，最终评测分数使用置信区间表达更为合适。为了简化置信区间的表达，本次榜单按 5 分一档划分区间，相同区间内的模型，可以近似认为属于同一水平。
- 以上所有测试结果，均基于各 Agent 产品的公开版本和网页入口，评测时所选择的 UI 界面模式和评测时间均标明在上述表格中。如果您发现模式选择有误，或者评测时间不是最新版本，或是自测分数与榜单分数存在显著差异，请及时联系我们进行勘误和重新评测，我们会尽快更新测试数据和榜单信息。
- 详细测评方法及关于 xbench-DeepSearch 的更多信息，请参考 Eval Card xbench-DeepSearch

xbench-DeepSearch Testing Score Distribution

- 本次 xbench 总计测评了 11 家公司的 21 个不同的产品，产品测试分数分布如下：

