Eval Card xbench-DeepSearch

A contamination free benchmark that measures the deep search ability of AI agents.

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。

例题

题目 1

问题	截至 2025 年 5 月 16 日，上交所科创板自开板以来，在所有注册生效状态且申报历时小于等于 180 天的公司中，多少家注册地是在中国沿海地区的？
答案	86 家
参考步骤	1. 搜索上交所发行上市项目审核动态：https://www.sse.com.cn/listing/renewal/ipo/ 2. 筛选审核状态：注册生效 3. 筛选更新日期-受理日期小于或者等于 180 天的公司 4. 搜索中国沿海省份或者直辖市的范围： 中国沿海地区通常包括以下省级行政区（11 个）： 直辖市：上海、天津 省份：辽宁、河北、山东、江苏、浙江、福建、广东、海南 自治区：广西壮族自治区 5. 符合中国沿海省份或者直辖市范围的公司数量计数：86 家

题目 2	
问题	黑龙江、吉林、辽宁，共有多少个地市级行政单位与外国接壤？
答案	12 个
参考步骤	参考步骤： 1. 搜索辽宁行政区划，确定只有丹东与朝鲜接壤。 2. 搜索吉林行政区划，确定只有延边州、通化市、白城市与朝鲜、俄罗斯接壤 3. 搜索黑龙江行政区划，牡丹江市、鸡西市、佳木斯市、鹤岗市、黑河市、双鸭山市、伊春市、大兴安岭地区，8 个地区与俄罗斯接壤。 4. 8+3+1=12

题目 3	
问题	《乐队的夏天》各季 top5 乐队中一共有多少名女性成员？
答案	6 名
参考步骤	1. 确定乐夏有三季 2. 确定每季 top5 乐队名字 3. 搜索每支乐队成员，找出其中女性成员 第一季：赵梦（新裤子）、石璐（刺猬） 第二季：刘敏（重塑雕像的权利） 第三季：冯海宁（Nova Heart）、其其格玛（安达组合）、赛汗尼亚（安达组合）

题目 4	
问题	尼米兹级航母第一艘下水到最后一艘下水期间，在任的美国总统有海军服役经历的平均服役时间是多少？ 备注 1：服役时间不计算预备役 备注 2：服役时间按结束减去开始年份计算，例如 1942 到 1943 算作 1 年
答案	4.5 年
参考步骤	参考步骤： 1、搜索 wiki，尼米兹级是第一艘尼米兹号 1972/05/13 下水，最后一艘乔治布什号 2006/10/09 下水 2、搜索美国总统就职时间，共有 7 位，从尼克松到小布什 3、搜索服役经历，其中海军 4 位，陆军 1 位，空军 1 位 4、计算服役年限：尼克松 4 年，福特 4 年，卡特 7 年，老布什 3 年 5、计算平均年限：4.5 年

题目 5	
问题	有一个被剪做鞋样的历史文物，对研究唐代均田制起到了重要的作用，这个文物中记载的年份，有一位唐朝的一代名相去世，请问这位名相有几个儿子？
答案	4 个
参考步骤	参考步骤： 1. 识别出这个历史文物是“赵怀满夏田契” 2. 识别出文物中记载的年份是贞观十七年（公元 643 年） 3. 识别出同年去世的名相叫做魏徵 4. 找到百度百科，计算出他有四个儿子

题目 6	
问题	根据中央音乐学院校外音乐水平考级细则，如果选用《新编中央音乐学院校外音乐水平考级教程丛书——钢琴（业余）考级教程》，从一级到演奏级共有十个级别，其音阶速度最低的平均值为♩=多少？将结果四舍五入到个位数。
答案	82
参考步骤	参考步骤： 1. 前往搜索网站，输入“中央音乐学院校外音乐水平考级细则” 2. 查询并找到中央音乐学院考级委员会官方网站 https://www.kaoji.com/#/ 3. 浏览并找到中央音乐学院校外音乐水平考级各专业考级细则，点击

	4. 浏览找到键盘类-钢琴-制定教程 1《新编中央音乐学院校外音乐水平考级教程丛书——钢琴（业余）考级教程》 5. 找到对应的基本练习考级要求 6. 提取一级至九级、演奏级音阶速度最低值，分别为 54，63，72，66，72，84，88，100，108，112 7. 计算音阶速度最低的平均值为 (54+63+72+66+72+84+88+100+108+112)÷10=81.9 8. 四舍五入到个位数为 82

题目 7	
问题	欧冠决赛历史上，分差最大的逆转中，获胜球员最后一位进球的是哪位球员？
答案	弗拉基米尔·斯米切尔
参考步骤	参考步骤： 1. 欧冠历史上，分数最大的逆转是利物浦和 AC 米兰，比分是 0:3 到 3:3 2. 常规时间，最后一位进球的是哈维·阿隆索在第 60 分钟的进球。 3. 但是比赛平分，没有结束，后续进入了加时和点球大战。因此点球大战还有进球 4. 最终点球大战利物浦 3:2 获胜夺冠 5. 利物浦罚点球的球员顺序是：迪特马尔·哈曼（◯），吉布里尔·西塞（◯），约翰·阿恩·里瑟（×），弗拉基米尔·斯米切尔（◯） 6. 最后一位是：弗拉基米尔·斯米切尔

题目 8	
问题	国家 A 是位于欧洲心脏地带的工业强国，2025 年 5 月国家 A 与东南亚国家 B 签署防务协议扩大双边合作。国家 B 的国果是什么？
答案	芒果
参考步骤	参考步骤： 1. 查询欧洲心脏地带国家，锁定工业强国。 2. 根据报道 https://www.zaobao.com.sg/news/sea/story20250515-6373031 2025 年 5 月德国与菲律宾签订防务协议扩大双边合作。 3. 菲律宾的“国果”是芒果。

题目 9	
问题	有一部短暂获得过奥斯卡最佳影片的电影，在最广为人知的一张海报上左手的手势被称作叫什么（英文）？

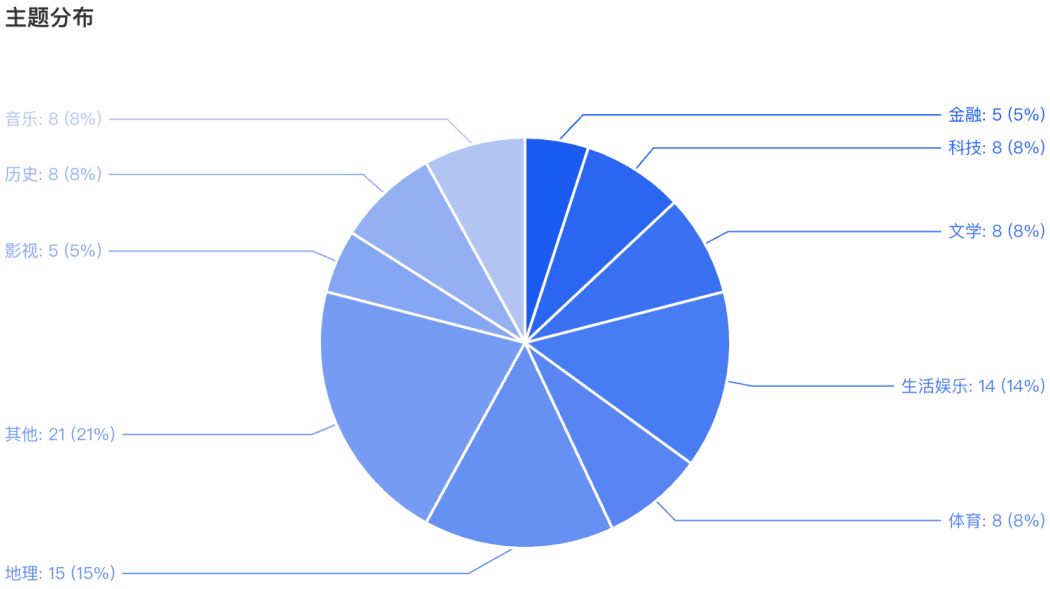
答案	Hamburger hands / La La Hand
参考步骤	1.分析“短暂获得” 推出电影指的是经历了颁奖乌龙的《爱乐之城》 2.检索其最知名的海报图像 3.分析海报中男主角左手的动作 为“hamburger hands”

题目 10	
问题	我在某天乘坐北京地铁 13 号线，出站后看到了马路对面有一间驾校，走了几百米后来到剧场看了一场话剧，印象深刻的台词和视力的变化有关。看完后下楼的过程中我看到的第一家餐厅和这片区域很相关，这家餐厅叫什么？
答案	东直门涮肉
参考步骤	1. 枚举北京地铁 13 号线沿线出站口有驾校的（不是） 2. 通过台词信息模糊定位到话剧《恋爱的犀牛》 3. 结合地铁+驾校信息，确定地点为东直门附近的蜂巢剧场 4. 枚举蜂巢剧场附近的餐厅 5. 找到符合所有条件的“东直门涮肉”

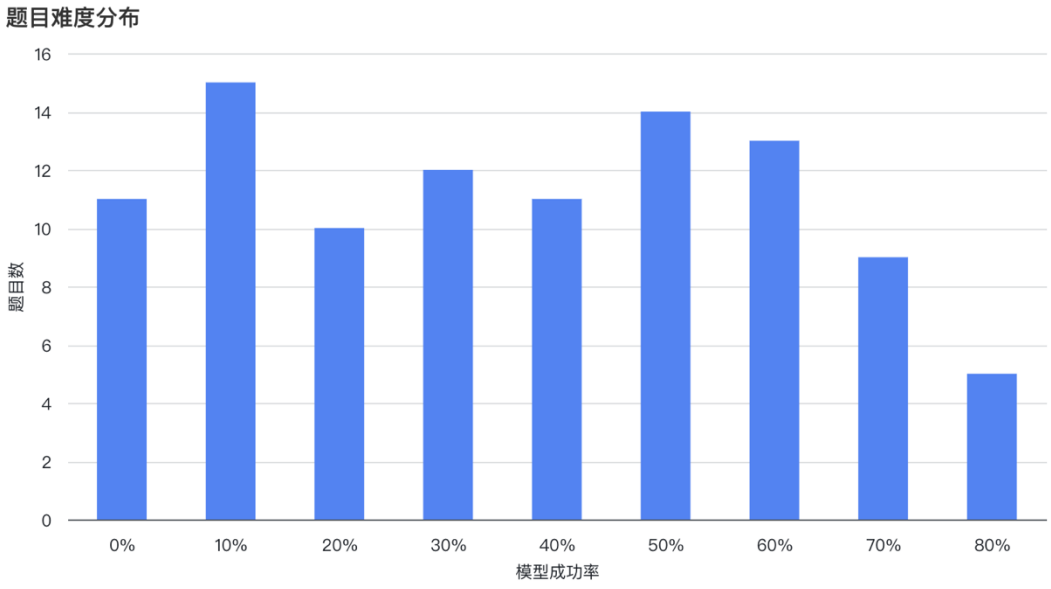
评估集详情

主题和难度分布

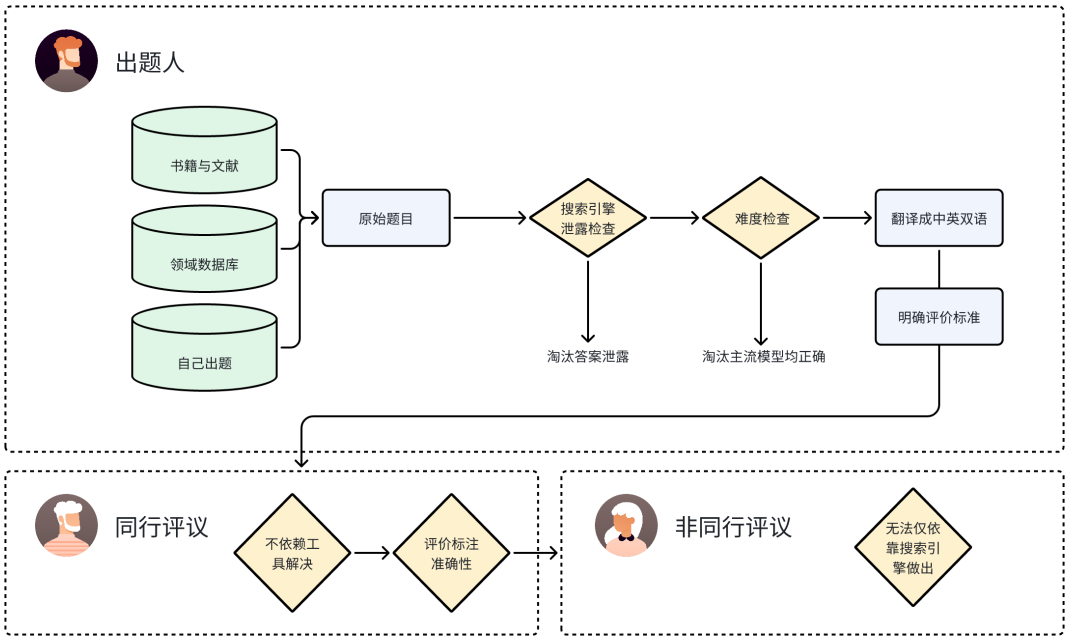
- 主题分布
    - 为了保证题目主题类型的多样性，我们参考了 OpenAI BrowseComp 评测集的分类，鼓励出题者围绕他们喜好的主题出题，不仅能够提高题目的质量和准确性，也帮助每个分类获得足够的题目覆盖。最终的题目类型分布见下图：



- 难度分布
  - 根据所测 30 组模型/Agents 的准确率结果，可以计算出每道题的模型成功率。该成功率可以作为题目难度的估计。如下图中，横轴是题目的模型成功率，纵轴是题目的数量。
  - 另一种难度估计方法，是统计真人解决问题所花的时间。该方法更接近人类对难度的感知，但缺点在于对评测同学领域知识的要求较高，一个行业内的同学和行业外的同学在解决问题时所花的时间差距很大，容易导致错误估计。我们计划在下次评测中给出真人所需耗时，作为难度分布的参考。



题目构建方法



- 在题目构建过程中，我们邀请了来自各行各业的 30 多位专家志愿者，依照我们给定的标注手册(见附录)进行出题。所出题目要求使用搜索引擎进行验证，保证不存在原题或是直接能够检索出答案。
- 为了保证题目具备难度和区分度，所有题目均需要经过主流模型的测试验证，淘汰正确率>80%的题目。
- 深度搜索的难题，一般涉及搜索空间大，或者推理的步骤多。因此出题者在出题时，会被指引尽量提供满足这两个条件之一的题目，以增加题目难度。
- 在部分出题过程中，我们参考了 BrowseComp 中使用的“想谜底，出谜面”的思路，鼓励出题者先根据给定的主题，随机想一个可以验证的事实谜底，然后根据谜底设计谜面。

> 示例：出一道考察搜索广度的题目
>
> 步骤一：先确定谜底为两位诺贝尔奖获得者大卫·贝克（David Baker）和大卫·维因兰德（David Jeffrey Wineland）
>
> 步骤二：设计有限的限制条件，引导模型在一个合理的搜索空间内进行深度搜索。这两位诺贝尔奖获得者，一位获得了物理学奖，一位获得了化学奖；两位都曾就职于华盛顿大学；最后加上两者的出生日期差别以保证答案的唯一性，这样一道搜索广度的题目就构建完成。
>
> 最终构建的题目为：一位诺贝尔物理学奖得主同一位诺贝尔奖化学奖得主的年龄相差 6799 天，他们两位有相同的 first name，曾就职于同一所位于美国西岸的大学，请问这两位诺贝尔奖得主是谁？

- 反之，另一种出题方式，是先想出谜面，逐步增加谜面的复杂度，最后设计出谜底来增加推理的深度：

> 示例：出一道考察推理深度的题目
>
> 步骤一：先确定一个出题者感兴趣的主题，如一件历史文物“赵怀满夏田契”
>
> 步骤二：为了考察推理深度，可以设计多层递进的条件。这件文物中记载了一个年份贞观十七年（公元 643 年），然后搜索该年份有什么重大的历史事件，可以搜到唐朝的名相魏徵去世，然后搜索魏徵，找到关于他的一个小众的事实点进行考察。
>
> 最终构建的题目为：有一个被剪做鞋样的历史文物，对研究唐代均田制起到了重要的作用，这个文物中记载的年份，有一位唐朝的一代名相去世，请问这位名相有几个儿子？

答案验证和质检

- 在标注者文档中，我们要求出题者提供详细的解题思路和步骤，并附上相关的信源链接。信源需要是权威性高的网站或者内容，例如政府网站，企业官方，主流媒体报告，wiki 百科等。来自于自媒体或二手新闻的信源一般不予以采纳。
- 与此同时，每道题都安排了两位博士生作为独立的质检员，先不参考解题思路自行尝试进行解答，并记录解答用时，最后对比答案。如果题目过难，质检员无法在 15 分钟内解答，或是题目完全超出了质检员的专业能力范围，则参考解题思路验证出题者给定的参考答案是否正确且唯一。
- 所有通过人工质检的题目，还需要验证主流模型的通过率：
    - 超过 80%主流模型能够通过的题目，会被舍弃。
    - 如果只有不到 10%的模型答对，或者大部分模型都回答了另一个答案，则提示答案可能潜在有误。题目和答案会被返回出题者进行二次校验。

评分方法

- 由于 Agent 产品是 UI 交互，没有 API 可以调用，因此我们组织了志愿者，将题目手工录入到各 Agent 产品中获取回复，并将收集到的回复由 LLM as Judge 统一进行评分。评分 prompt 请参考本文附录。

已知问题

- 难度分布：虽然大部分主流产品的得分不足 50%，但 SOTA 模型如 o3 联网模式可以获得接近 70%的通过率，表明本评测集对于最先进的模型来说，难度仍然偏低。我们会持续收集优化题目难度。
- 题目数量：由于深度搜索产品普遍需要较长的搜索解答时长，一般需要 5～30 分钟，且没有 API 能够自动获取回复，需要较高的成本进行人工手动评测，因此本次评测只保留了 100 题作为初版题库。对于能够使用自身内部 API 进行自动评测的 Agent 开发者来说，如果有更多的题量，可以获得更加置信的结果。为此，xbench-DeepSearch 会持续发布新的高质量的题目，提供更加置信的评测结果。
- 题目勘误：涉及专业知识的题目，受限于质检员自身的知识储备，无法保证所有题目没有歧义，答案没有瑕疵。如果您在使用本题库的过程中，发现题目有不清晰/歧义/错误答案等情况，欢迎及时联系 team@xbench.org，我们会尽快勘正。

联系方式

- 如果您是评测爱好者，想要参与到评测集的建设中；
- 如果您是 Agent 开发者，想要提交您的产品参与评测，或者提交白盒分数；
- 如果您希望给我们反馈意见，对题目或评估结果有疑问；
- 欢迎联系 team@xbench.org，我们会尽快反馈。

附录

标注者文档

收集新题目的数据集（不是已有问题的变体），要求：

- 确保您的问题基于可靠来源（Arxiv、百度百科、Bilibili、国内外政府与国际组织网站）。对于 2 级和 3 级问题，创建问题的好方法是结合多个可靠来源。
- 确保您问题的答案在互联网上不以纯文本形式存在。
- 确保您问题的答案是一个数字或最多几个词，以使评估更加可靠。
- 确保您问题的答案不会随时间变化。这包括可靠来源可能被删除的情况。所需搜索的网页时间必须在 2024 年 12 月 31 日之前。
- 确保您问题的答案是明确无歧义的。
- 确保您的问题是"有趣的"，即阅读后您认为 AI 助手能够回答这类问题会对您有很大帮助。
- 确保人类标注者能在合理时间内回答您的问题。
- 检查包含回答所需信息的网站的 robots.txt 文件，确保 AI 助手可以访问。如果一个网页以及出现在搜索引擎的结果中，则该网站没有问题。

题目收集表格样例

问题代号	ASE-标注者编号[xxxxxx]-语言[ZH/EN]-难度编号[L1/L2/L3]-问题编号[xxxxxx]
难度级别	1/2/3
问题	问题表述

领域	xxx
答案	判题答案
解题步骤	1. ... 2. ... 3. ... 4. ... 可以校验地操作步骤
步骤数量	4
解题时间	预期花费的时间
备注	额外补充的特殊信息

LLM as Judge prompt

我们使用了 Humanity’s Last Exam (Phan et al., 2025) 中同样的 judge prompt 作为自动评分方法。

```
python
JUDGE_PROMPT = """Judge whether the following [response] to [question] is correct or not based on the precise and unambiguous [correct_answer] below. [question]: {question} [response]: {response} Your judgement must be in the format and criteria specified below: extracted_final_answer: The final exact answer extracted from the [response]. Put the extracted answer as ’None’ if there is no exact, final answer to extract from the response. [correct_answer]: {correct_answer} reasoning: Explain why the extracted_final_answer is correct or incorrect based on [correct_answer], focusing only on if there are meaningful differences between [correct_answer] and the extracted_final_answer. Do not comment on any background to the problem, do not attempt to solve the problem, do not argue for any answer different than [correct_answer], focus only on whether the answers match. correct: Answer ’yes’ if extracted_final_answer
```

matches the [correct_answer] given above, or is within a small margin of error for numerical problems. Answer ’ no’ otherwise, i.e. if there if there is any inconsistency, ambiguity, non-equivalency, or if the extracted answer is incorrect. confidence: The extracted confidence score between 0|\%| and 100|\%| from [response]. Put 100 if there is no confidence score available."""