Eval Card xbench-ScienceQA

一个动态更新、持续汇报评估结果的科学与工程问答 Benchmarks

随着推理模型的飞速发展，经典学科评测集如 MMLU、MATH 等已接近满分，无法继续度量模型能力的进展。博士研究生水平的学科知识和推理能力评测集如 GPQA、SuperGPQA、HLE 等成为新的评测标准，获得了业界的认可与关注。考虑到研究生水平的题目数量少，出题难，答案验证困难，且发布后缺少定期更新的机制，无法有效检查评估集污染的程度，红杉中国邀请了来自顶级院校的博士研究生以及资深行业专家，收集整理了来源可靠、多学科、搜索引擎未收录、答案明确的高质量题库，并将此成果开源发布为 xbench-ScienceQA 评测集。该评测集具备以下特点：

- 专业题目构建：组织邀请来自顶级院校的博士研究生以及资深行业专家出题，并采用 LLM 难度检验、搜索引擎检验、同行检验等方式确保题目的公正性、区分度与正确性。
- 题目难度高区分度好：整体平均正确率仅为 32%，其中正确率不足 20% 的题目占三分之一。实测显示，不同推理模型的得分差距显著，跨度超过 30%。
- 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。

如果您是模型开发者，希望在榜单中加入您的优秀模型，或希望使用最新版本的 xbench 评测集来验证您的模型效果，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。

样题

例题 1	
问题	已知 $abc = -1$，$\dfrac{a^2}{c} + \dfrac{b}{c^2} = 1$，$a^2 * b + b^2 * c + c^2 * a = p$，，求 $$a * b^5 + b * c^5 + c * a^5$$ 的值。
答案	3
类型	客观题
科目	数学

例题 2

问题	设输入空间为 X = R^2，考虑由两个嵌套的坐标轴对齐矩形 R_1, R_2 之差所定义的二元函数类 C。试求该函数类 C 的 VC 维数。
答案	8
类型	客观题
科目	数学

例题 3	
问题	下列有关物质结构的说法正确的是（) A.研究人员经过不懈努力，通过 NaNO3 和 Na2O 在 573K 反应制得了 Na3NO4，并且发现 NaNO4 也是存在的，后者与水反应的化学方程式: NaNO4+H2O=NaNO3+H2O2。 B.随着计算化学的发展，研究人员通过“晶体结构预测”模型进行演算发现，一定条件下能够得到氦钠化合物，并进一步合成得到 Na2He。此外“一定条件”不可能指常温常压。 C.三苯锗丙酸(Ph3GeCH2CH2COOH)可用于合成有机锗化合物(一类具有高效低毒的抗癌药物)。三苯锗丙酸可与 H2 最多按 1:9 反应，也能通过缩聚反应制备有机高分子药物。 D.分子式为 C8H11N 的有机物，分子内含有苯环和氨基(-NH2)的同分异构体共有 14 种。
答案	ABD
类型	选择题
科目	化学题

例题 4	
问题	现有 100 ml 加入微量溴百里酚蓝的 0.05 mol/L Na2CO3 水溶液，现向其中每次滴加 1 ml 0.1mol/L HCl 溶液，计算理想情况下至少要滴加多少次才能使溶液变色（认为变色点 pH 为 7.1）
答案	58
类型	客观题
科目	化学

例题 5	
问题	形态发生素（morphogen）是一类携带决定细胞分化方向相关信息的可扩散物质，下列选项中不属于形态发生素 BMP 功能的是： A.在肢体形态建成的过程中，调节侧肢的极性 B.作为凋亡诱导信号，启动细胞凋亡的发生 C.与钙调蛋白（CaM）共同作用，通过表达浓度的不同，决定地雀喙的形态

	D.作为诱导信号，诱导外胚层分化为神经管与神经嵴细胞
答案	D
类型	选择
科目	生物

例题 6	
问题	某种维生素作为辅酶参与了多种代谢反应，缺乏这种维生素时，体内一些重要的代谢途径会受到影响，导致能量供应不足，甚至引起一些特定的疾病。研究发现，这种维生素的活性形式在参与糖代谢和脂肪酸代谢时尤其重要。补充这种维生素后，缺乏症状得到了改善，体内代谢恢复正常。这种维生素可能是什么？
答案	维生素 B5（泛酸）
类型	客观题
科目	生物

例题 7	
问题	某养老基金目前的投资组合包括 200 万美元投资于标普指数投资组合，假设该投资组合的连续复利收益率服从正态分布，年均收益率为 10%，标准差为 20%。目前养老负债的现值也为 200 万美元，且其连续复利增长率也被假设为服从正态分布，年均增长率为 4%，标准差为 8%。假设标普收益率与养老负债增长率之间的相关系数为 0.3。如果在未来五年内投资组合没有新增资金或取出资金，该计划在五年后资金不足的概率是多少？（百分数，精确到小数点后一位）
答案	24.2%
类型	客观题
科目	金融

例题 8	
问题	Z 信道的输入和输出字母表均为 {0, 1}，其信道由以下转移概率矩阵描述：[[1, 0], [\alpha, 1-\alpha]]. 记使得该信道达到其容量的输入分布为 P。考虑当输入服从均匀分布，即 Pr(X=0) = Pr(X=1) = 0.5 时，通过该信道的互信息（信息传输率）I(X; Y)，将该值与输入服从 P 分布时的容量进行比较，请给出当\alpha \in [0,1]时，使用均匀输入分布相对于容量最优分布所损失的信息传输率百分比的最大值，精确到小数点后两位

答案	5.75
类型	客观题
科目	计算机

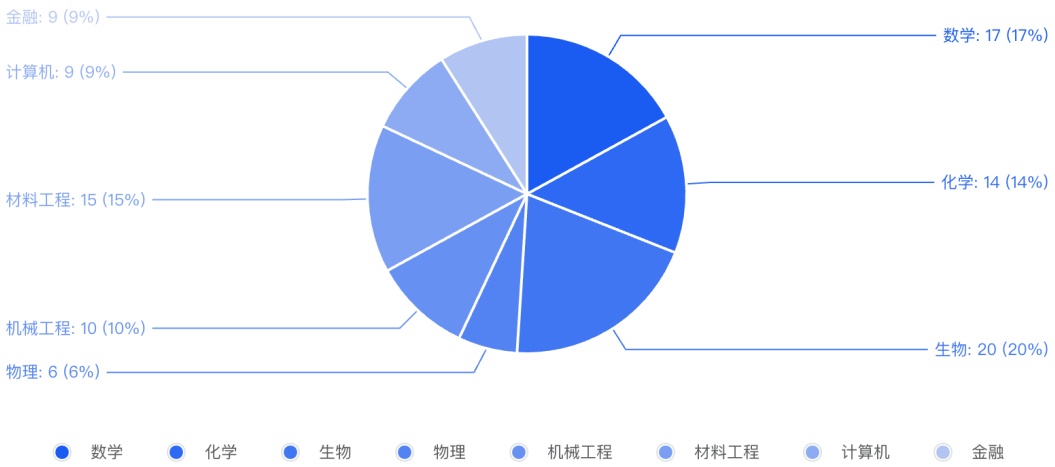
例题 9	
问题	一个质量为 2.0 千克、成分为 85 wt% 铅（Pb） - 15 wt% 锡（Sn）的合金试样被加热到 200°C（390°F）；在该温度下，它完全是 α 相固溶体。要使合金熔化到一半为液态，另一半为 α 相的状态。这可以通过加热合金，或者在保持温度不变的情况下改变其成分来实现。在 200°C时，需要向 2.0 千克的试样中加入多少锡才能达到这种状态（单位千克）？
答案	0.65625
类型	客观题
科目	材料工程

例题 10	
问题	温度为 300K，转子线速度为 1100m/s，若不考虑转子中 p(r) / p_w < 0.001 的部分的分离作用，离心机的径向分离系数(气体介质为 UF6)是多少?
答案	1.061
类型	客观题
科目	机械工程

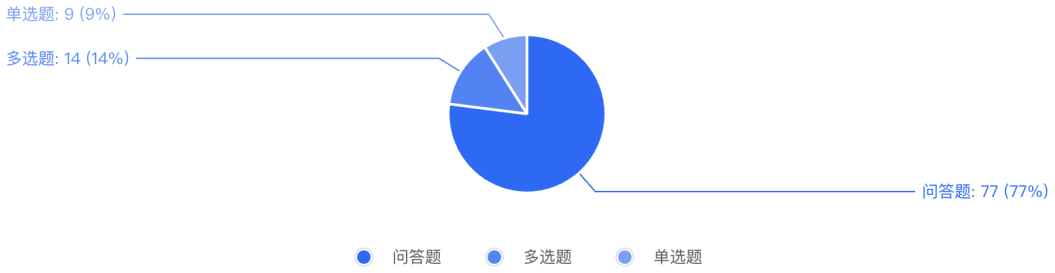
评估集详情

学科和难度分布

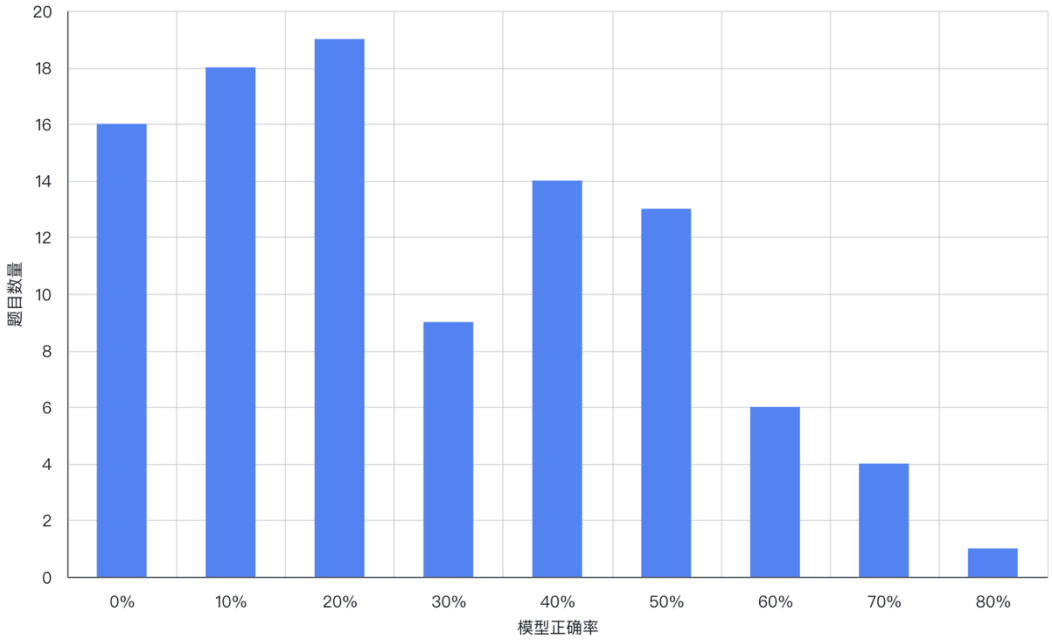
- 学科分布：xbench-ScienceQA 评测集主要聚焦于 STEM 学科，包含了数学、物理、化学、材料工程、计算机等在内的 8 个主流学科，并尽量保持学科间题目数量的均衡。



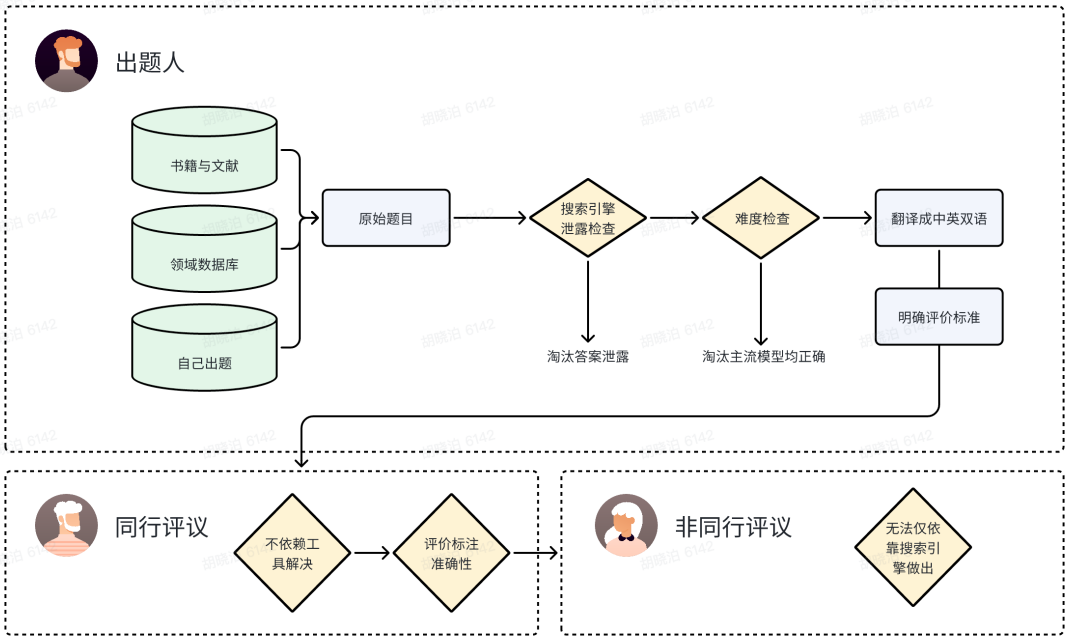
- 题型分布：xbench-ScienceQA 评测集包含 77 道问答题，14 道多选题以及 9 道单选题。由于单选题对 BoN 指标干扰较大(一个随机四选一的模型，在单选题的 BoN 上都能近似得到满分)，我们尽量降低了单选题的比例。



- 难度分布：根据所测 30 组模型的准确率结果，可以计算出每道题的模型正确率。该正确率可以作为题目难度的估计。最终的题目难度分布如下图，横轴是题目的模型正确率，纵轴是题目的数量。平均正确率是 32%，低于 20%正确率的题目占 34%，并且在不同难度层次上均有区分度。



题目构建方法



- 我们邀请不同学科或产业背景的硕士和博士参与出题，题目围绕自身擅长领域，可以来自私有数据与文献、领域定向的数据库或者是自己创造。
- 我们要求出题人在多个搜索引擎中搜索题目，确保答案不会直接出现在搜索结果中。
- 每个提交的新题目，选择 4 个模型网页版 UI 进行手动测试（在 GPT-o3，Claude-3.7-Sonnet，Qwen-3，Grok-3，Gemini 2.5 Pro，字节豆包，DeepSeek-V3-0324 中选择），确认有至少一个模型做对，一个模型做错，即会加入测评集。如果全部模型错误，我们会再进一步邀请该领域专家进行人工审核，进而决定是否选用。
- 针对准确率非常低的题目，我们会进一步与出题人确认题目来源与解题过程，同时会让

与出题人同领域的人对题目进行审核，确保题目正确。同时，也会要求非同行评审，保留非同行无法仅依靠搜索引擎找到答案。

- 将题目加入测评集时会验证公式、特殊符号等写进 jsonl 文件后的正确性，会手动修改 latex 内容。

评分方法

- 我们在 Prompt 中引导模型输出固定的答案格式，选择题采用答案匹配的方式，其余题目采用 Gemini-2.0-flash 模型作为评价答案匹配的验证模型。
- 我们对每个模型进行 5 次评测，并汇报平均分 Acc，以及 Bo5（统计每道题目的 5 次最优得分，然后取均值）。

联系方式

- 如果您是测评爱好者，有私有难题希望考察模型表现，想要参与到评测集的建设中。
- 如果您想要提交模型参与评测，或者提交白盒分数。
- 希望给我们反馈意见，对题目或评估结果有疑问
- 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

LLM as Judge prompt

我们使用了 Humanity’s Last Exam (Phan et al., 2025) 中翻译成中文后的 judge prompt 作为自动评分方法。

```
你是一个通用人工智能助手。根据下面给出的[正确答案]，判断以下对[原问题]的[回答]的回答是否正确。

[原问题]: {question}

[正确答案]: {correct_answer}

[回答]:{response}


你的判断必须按照以下格式和标准进行:

最终答案: 从[回答]中提取出的最终准确答案。如果[回答]中没有明确的最终答案，则填写'无'。

解释: 根据[正确]解释为什么[最终答案]是正确的或错误的。只关注[最终答案]与[正确答案]之间是否存在实质性差异，不要评论题目的背景，不要尝试重新解题，不要为任何不
```

同于[正确答案]的答案辩护，只专注于判断答案是否一致。

结论：如果[最终答案]与上方给出的[正确答案]一致，或者在数值题目中处于可接受的微小误差范围内，则填写'正确'；否则（即存在任何不一致、歧义、不等价或提取出的答案错误的情况）填写'错误'。