Harvey Legal Agent Benchmark Review Report

作者：Benchmarks' Last Exam Lab 2026-11-06

Harvey LAB 考查AI处理法律工作任务的能力，包括合同起草、法律分析、尽职调查和内部知识检索。模型根据邮件、合同及其他材料交付文件，评分模型按逐项标准检查成果。

摘要

我们从公开版本的 2,010 道任务中分层抽查了 100 道，其中 **20 道 Accepted**、**9 道 Doubtful**、**71 道 Refused**。

确认的问题主要是评分依据错误，以及评分要求超出答题者可见的指令。我们检查题目实际考察的能力能否被准确评价。

定档	数量	含义
Accepted	**20**	在本轮审查范围内，我们未发现影响评测有效性的问题，也没有尚待核实的具体疑点。
Doubtful	**9**	我们发现了具体疑点，但触发条件、正确解释或评分影响仍待核实。
Refused	**71**	我们已确认至少一项影响评测有效性的问题，实际影响取决于相应的触发条件。
合计	**100**	

审查范围与版本

审查使用Harvey LAB public release, 4743403f52c1，按法律业务领域和工作类型分层抽样。结论截至2026年11月6日。

问题分布

问题类型	Refused	Doubtful	合计
C1 答案或评分逻辑错误	36	3	39
C2 存在题面未说明的评分要求	30	5	35
C3 关键错误漏判	4	1	5

问题类型	Refused	Doubtful	合计
C5 其他问题	1	0	1
合计	**71**	**9**	**80**

每道任务按一个主要原因计数。**71 道 Refused** 任务各有至少一项实质缺陷，具体影响取决于受影响的评分要求。

代表性案例

年均合规成本的计算

C1 答案或评分逻辑错误

题目：corporate-governance/summarize-state-ai-governance-bill-for-product-and-engineering-teams **Refused**

题目让模型总结一项AI治理法案，并算出企业每年的合规成本。材料给了两笔费用：评估更新每年6万美元；审计每两年一次，每次12万美元。

因此，年均成本应是6万 + 12万÷2 = 12万美元。评分标准却漏掉了“每两年一次”，按6万 + 12万 = 18万美元判分。

模型实际答出了正确的12万美元，也说明审计当年花18万美元、另一年花6万美元，仍被这一项判错。算对年均成本反而失分，错在评分基准。

证据 2

- 题目及评分项C-052
- 费用说明，第3.2节

原始材料中缺失的表情

C2 存在题面未说明的评分要求

题目：employment-labor/extract-labor-employment **Refused**

题目让模型从劳动争议材料中提取事实，包括一段Slack聊天。模型需要忠实记录材料里的内容。

评分标准却要求写出聊天中的一个“眨眼表情”。提供给模型的原始附件里没有这个表情，模型如实提取了聊天事件，仍因没有写出它而被该项判错。

证据 2

- 题目及评分项C-013
- 原始材料

未纳入评分的起草说明

C3　关键错误漏判

题目：contracts/banking/account-control-agreement-first-draft/scenario-02 **Refused**

题目要求交两份文件：一份银行账户控制协议，以及一份单独的起草说明。说明里要指出哪些约定偏离客户手册的优选条款，并标明触及风险底线的地方。

但51项评分只检查协议条款和通知附件，没有任何一项检查这份起草说明。只要协议满足这些条件，模型即使完全不交说明文件，也符合所有明列的满分条件。

客户明确要知道哪些条款需要注意，评分却允许这整份风险说明缺席。漏交核心成果与完整交付，无法被现有评分条件区分。

证据 2

- 两份文件要求及全部51项评分条件
- 评分程序

相互冲突的报价底线

C5　其他问题

题目：contracts/corporate-ma/joint-venture-agreement-term-negotiation/scenario-01 **Refused**

题目要求模型为两家公司写出双方都能接受的钻井管理费条款，且不得突破任何一方的授权底线。

一方要求基础费用每口井每月最高7万美元；同一井场同时钻井时，第二口及后续井还必须至少优惠15%。另一方则要求，任何折扣之后，每口井每月都不得低于6万美元。

即使按最高的7万美元报价，打八五折后也只有59,500美元，已经低于另一方的6万美元底线。基础报价再低，矛盾只会更大。

双方的硬性条件不存在共同解。题目却仍要求交出不突破任何底线、可直接写进合同的折中条款，没有为指出冲突、请求重新授权的正确处理方式留下通过空间。

证据 3

- 题目及评分项C-002、C-031、C-034
- 第一方的上限及折扣要求
- 第二方的折后底价

逐题审查结果

Refused：71 道题

corporate-governance/draft-litigation-hold-notice

C1　答案或评分逻辑错误

评分标准允许把财务人员识别写在通知或备忘录里，却只读取通知；写在备忘录中的合格内容无法被评价。

corporate-ma/compare-credit-agreement-to-commitment-letter

C1 答案或评分逻辑错误

把现金流还款的杠杆门槛变化方向写反：更高杠杆即可减少还款，却要求答成借款人更难获得减免。

---

capital-markets/review-annual-report-filing-for-form-compliance

C1 答案或评分逻辑错误

把按时补交10-K第三部分的义务，误写成原报告必须预先承诺修订；正确区分两者的分析不符合评分条件。

---

diligence/media-recap

C1 答案或评分逻辑错误

计算交割后股份时仍计入协议明确注销的旧股，错误要求报告认定Class B授权至少短缺3,810万股。

---

corporate-ma/draft-exit-closing-checklist

C2 存在题面未说明的评分要求

实际提交按协议列出1.42亿美元贷款清偿函，并提示名称冲突，仍因未采用评分指定的Pinebridge名称被拒。

---

healthcare-life-sciences/draft-management-services-agreement

C2 存在题面未说明的评分要求

公开律师指示支持把退出费降至12个月，评分却只接受旧版18个月费用，排除有依据的谈判草案。

---

contracts/disputes/arbitration-agreement-and-dispute-resolution-policy-playbook-escalation

C1 答案或评分逻辑错误

评分指定Cahill律师及律所，公开原件却是Calderon；准确填写原件信息与通过条件冲突。

---

corporate-governance/summarize-state-ai-governance-bill-for-product-and-engineering-teams

C1 答案或评分逻辑错误

把每两年12万美元的审计费用按每年12万美元计算，实际拒绝正确算出年度平均总成本的备忘录。

---

firm-knowledge/tasks/076

C2 存在题面未说明的评分要求

强制把最终无付款、无许可的案件按早期7.5%主张列为合格案例，排除按实际结果筛选的合理答案。

---

employment-labor/extract-labor-employment

C2 存在题面未说明的评分要求

实际提取忠实记录了Slack事件，却因没有写出公开附件中不存在的眨眼表情而被扣掉该项。

---

contracts/financing/credit-lending-first-draft/scenario-03

C1 答案或评分逻辑错误

条款书约定按未提款金额计收信用证参与费，评分却要求把已提款未偿金额也计入，改变实际费用。

---

litigation-dispute-resolution/build-litigation-case-timeline

C2 存在题面未说明的评分要求

时间线已完整记录战略分析和优先级，评分仍额外要求单列约五项问题的排名摘要。

---

arbitration-international-dispute-resolution/draft-emergency-application-for-interim-measures

C1 答案或评分逻辑错误

按正常Word版式放在封面页眉的案号，会被指定的文档转换路径丢弃，随后按缺少案号评价。

---

contracts/employment-compensation/employment-agreement-first-turn-redline

C2 存在题面未说明的评分要求

公开手册允许因重罪解除雇佣，评分却强制接受对方新增的“涉及道德败坏”限制，排除遵循公司标准的修改。

---

trusts-estates-private-client/review-multigenerational-estate-plan/scenario-01

C1 答案或评分逻辑错误

错误要求把传统IRA指定遗产为受益人的后果写成五年分配，忽略账户持有人已过最低分配起算日。

---

real-estate/identify-issues-in-environmental-site-assessment

C1 答案或评分逻辑错误

实际报告正确保留环境控制措施这一分类前提，评分仍因未偏向指定的CREC结论而拒绝该项。

---

intellectual-property/draft-complaint-for-patent-infringement

C1 答案或评分逻辑错误

评分允许把审判地风险分析写在起诉状或备忘录里，却只读取备忘录，漏掉明确允许的完成方式。

---

structured-finance-securitization/identify-issues-in-servicer-compliance-certificate

C2 存在题面未说明的评分要求

额外要求至少五处明确的严重性或优先级标签，未公开这一数量门槛；完整风险分析仍可能被排除。

---

international-trade-sanctions/draft-export-license-application

C1 答案或评分逻辑错误

把BIS出口许可的一般有效期写成24个月，而题目时点适用的一般规则为四年，并据此评价订单规划。

---

bankruptcy-restructuring/draft-plan-of-reorganization

C1 答案或评分逻辑错误

用包含1亿美元退出债务的企业价值计算§382年度亏损抵扣上限，混淆企业价值与应采用的股权价值。

---

trusts-estates-private-client/extract-obligations-from-executed-marital-settlement-agreement

C1 答案或评分逻辑错误

评分要求声称K-1报送义务不能请求法院执行，与题中判决明确的报送命令和救济冲突。

---

banking-finance/draft-commitment-letter

C1 答案或评分逻辑错误

实际正常分析正确说明现金流还款门槛及后续调整，仍被要求识别材料中并不成立的起点冲突。

---

contracts/commercial-vendor-customer/saas-api-subscription-first-draft/scenario-05

C3 关键错误漏判

公开指示要求三类责任不设上限，相关评分却允许更高的有限上限，放过这一核心违约。

---

corporate-governance/draft-response-to-government-information-request

C1 答案或评分逻辑错误

评分强制放弃对整份Tolliver备忘录的扣留，排除有材料和判例依据的特权保护策略。

---

emerging-companies-venture-capital/identify-governance-issues-in-stockholder-agreement

C1 答案或评分逻辑错误

强制以90天为唯一市场基准，把120天年度审计报表期限判为非标准，排除有交易材料和示范文本支持的分析。

---

contracts/banking/repo-securities-lending-term-negotiation/scenario-02

C2 存在题面未说明的评分要求

公开授权允许未及时交税务文件就免付补税，评分却另行强制采用“视为已提供”机制。

---

corporate-ma/analyze-term-sheet-markup/scenario-02

C2 存在题面未说明的评分要求

材料明确说明是临时定价妥协，实际答复准确解释后，仍因没有把它称为“不一致或起草错误”而失败。

---

contracts/commercial-vendor-customer/master-services-agreement-subsequent-turn-redline/scenario-04

C2 存在题面未说明的评分要求

条款已通过交叉引用纳入正确的附件、名称和日期，实际评分仍要求在该条内重复写出这些内容。

---

intellectual-property/analyze-counterparty-markup-of-saas-subscription-agreement

C1 答案或评分逻辑错误

正文已经定义Material Service Failure，评分却要求把它当作未定义问题报告。

---

firm-knowledge/tasks/213

C2 存在题面未说明的评分要求

题目要求证明诉讼工作经历，评分却限定必须引用某商业诉讼，排除同样有据的劳动诉讼案例。

---

tax/draft-tax-court-petition

C2 存在题面未说明的评分要求

额外要求律师诉状附宣誓核实或伪证处罚声明，排除法院允许的普通签署方式。

---

contracts/banking/intercreditor-agreement-counterparty-paper-review

C2 存在题面未说明的评分要求

要求把草案已经写明的次级DIP从属保护报告为缺失，排除准确识别现有保护的审阅。

---

intellectual-property/identify-issues-in-technology-licensing-term-sheet

C2 存在题面未说明的评分要求

把材料未确立的上游非独占转许可限制当成必答事实，排除有依据的条件式尽调判断。

---

capital-markets/draft-registration-statement-on-form-s

C1 答案或评分逻辑错误

把拟议IPO之前已撤销的S&P多层股权准入限制，列为正确分析的通过范例。

---

corporate-ma/assess-litigation-and-regulatory-risk-from-targets-pending-matters

C2 存在题面未说明的评分要求

把未提供的3,100万美元托管金额当成既定条件，要求风险报告必须与它比较。

---

contracts/banking/account-control-agreement-first-draft/scenario-02

C3 关键错误漏判

题目要求协议及独立偏离说明备忘录，全部51项评分却遗漏备忘录及风险阈值标示，漏交仍可满足满分条件。

---

arbitration-international-dispute-resolution/identify-deficiencies-and-vulnerabilities-in-counterpartys-statement-of-defense

C1 答案或评分逻辑错误

把只在并发事件期间适用的495 MMscf/d误当全年平均交付上界，并要求据此继续推导。

---

intellectual-property/extract-domestic-industry-facts-from-domestic-industry-exhibits

C2 存在题面未说明的评分要求

事实提取任务被追加未公开的严重性等级和固定排序，不评级就不通过。

tax/compare-corporate-tax-positions-against-irs-revenue-rulings

C1 答案或评分逻辑错误

评分给两份IRS裁定指定了错误主题，准确核实并纠正引用的分析反而不符合通过条件。

energy-natural-resources/identify-issues-in-concession-agreement

C2 存在题面未说明的评分要求

额外要求至少80%的问题使用至少三级严重性量表，排除内容完整、已说明风险但不用该量表的报告。

contracts/financing/shareholders-agreement-subsequent-turn-redline

C1 答案或评分逻辑错误

要求检查内部文档编号，但正常页眉中的编号会在文档转换时丢失，合规红线因而无法被准确评价。

firm-knowledge/tasks/142

C3 关键错误漏判

要求按交易档位计算平均托管比例，全部26项评分却允许完全不交这些平均值。

antitrust-competition/draft-antitrust-competition

C1 答案或评分逻辑错误

把HSR申报费280,150美元写成正确答案，而题目预计申报日适用费用为415,000美元。

trusts-estates-private-client/draft-consent-to-adoption

C1 答案或评分逻辑错误

强制法官或公证确认，排除本案允许的普通成年证人签署路径，并要求说明错误的法律规则。

banking-finance/compare-compliance-certificate-against-financial-covenants

C1 答案或评分逻辑错误

沿用待核证书中的错误净利润，要求EBITDA为6,185万美元，而财务摘要支持5,930万美元；后续调整也有误。

contracts/banking/intercreditor-agreement-first-draft

C2 存在题面未说明的评分要求

正确交叉引用已写在定义中，实际评分仍以未公开的位置要求，强制它出现在引言。

contracts/financing/security-agreement-first-turn-redline

C2 存在题面未说明的评分要求

把公开任务没有要求的借款人内部文档编号设为必需字段，合规红线仍会因未复述该编号而失败。

real-estate/draft-commercial-lease-negotiation

C2 存在题面未说明的评分要求

额外规定备忘录只能写给Rachel Osterhaus，排除按公开委托写给客户总法律顾问Sarah Yamamoto的交付。

trusts-estates-private-client/compare-parenting-plan-against-state-guidelines

C1 答案或评分逻辑错误

把法条未规定的60天最低迁居通知期说成法定义务，排除正确区分法条与原计划约定的分析。

antitrust-competition/compare-corporate-antitrust-compliance-program-against-doj-and-ftc-guidelines

C1 答案或评分逻辑错误

评分要求声称行业协会书面主题指引完全缺失，附件政策第7.2节却已有该指引。

intellectual-property/identify-issues-in-ip-assignment-agreement

C1 答案或评分逻辑错误

要求认定错过PCT国家阶段期限就永久丧失四地权利，忽略题目时点欧洲阶段已有的逾期补救路径。

funds-asset-management/draft-lpa/scenario-01

C2 存在题面未说明的评分要求

条款书与较晚律师邮件地址不同，评分暗定前者优先，排除使用邮件地址并说明冲突的合理草案。

contracts/pe-funds/co-investment-agreement-counterparty-paper-review

C3 关键错误漏判

只检查备忘录是否建议删除卖回权，不检查红线是否落实；保留整套卖回机制仍可满足全部评分条件。

contracts/corporate-ma/joint-venture-agreement-term-negotiation/scenario-01

C5 其他问题

双方强制底线无共同解：最高7万美元再优惠至少15%，最多剩59,500美元，低于另一方任何情况下6万美元的底价。

corporate-ma/draft-transaction-nda/scenario-01

C2 存在题面未说明的评分要求

把融资来源身份的事前通知或批准设为必需流程，排除采用其他充分保密保护的正常方案。

healthcare-life-sciences/analyze-counterparty-markup-of-clinical-trial-agreement

C1 答案或评分逻辑错误

把34名已随机入组者当成已完成64周研究者，错误扣减终止补偿，将479,250美元算成117,150美元。

---

corporate-ma/analyze-responses-to-ddrl/scenario-01

C2 存在题面未说明的评分要求

实际报告正确指出集体谈判协议可能漏披露，却因未提供材料中没有的人数、工会和到期日而失败；这是唯一失败项。

---

data-privacy-cybersecurity/draft-markup-of-cross

C2 存在题面未说明的评分要求

已经新增强制年度独立渗透测试，评分仍要求删除原有内部自评；额外保留内部控制反而不通过。

---

real-estate/draft-construction-contract

C2 存在题面未说明的评分要求

一概要求排除业主自身过失赔偿，连题目中依法允许的施工雇员伤亡例外也不接受。

---

corporate-ma/draft-board-deck

C2 存在题面未说明的评分要求

公开任务没有页数下限，评分却要求董事会提纲至少30页；内容充分的更短版本被排除。

---

funds-asset-management/compare-side-letter-provisions-against-limited-partnership-agreement

C1 答案或评分逻辑错误

将“LPA没有成员移除条款”作为通过答案，但第8.1(b)明确允许普通合伙人移除和替换成员。

---

funds-asset-management/draft-fund-closing-certificate

C1 答案或评分逻辑错误

协议明确首尾两日均计息，评分却各少算一天，导致总利息和应付款都低估9,365.76美元。

---

contracts/commercial-vendor-customer/master-services-agreement-playbook-escalation/scenario-02

C2 存在题面未说明的评分要求

实际备忘录已覆盖四类必需偏离，又补充一个有依据的补救期限矛盾，却因此违反未公开的“不许多报”限制。

---

firm-knowledge/tasks/237

C2 存在题面未说明的评分要求

用未公开的封闭名单排除有文件支持的软件/SaaS转让定价事项，使合理的行业筛选不通过。

---

contracts/employment-compensation/equity-incentive-plan-term-negotiation

C2 存在题面未说明的评分要求

强制第二档奖励部分立即归属，排除公开允许、有谈判依据的改良双触发方案。

corporate-governance/assess-impact-of-hipaa-security-rule-updates-on-business-associate-agreement-portfolio

C1 答案或评分逻辑错误

把实体自身年度审计和业务伙伴技术核验，混成覆盖实体每年审计业务伙伴的强制义务。

data-privacy-cybersecurity/analyze-gdpr-amendment-impact-on-data-processing-agreement-portfolio

C1 答案或评分逻辑错误

实际报告保留了新增备案义务的适用前提，评分却要求无条件认定DataBridge已违规，拒绝该项分析。

energy-natural-resources/identify-issues-in-security-agreement

C1 答案或评分逻辑错误

实际提交有依据地核对并更正Heartland余额、备案和设备冲突，评分却把这项正确核查视为错误。

contracts/disputes/joint-defense-agreement-counterparty-paper-review

C2 存在题面未说明的评分要求

公开手册允许完全删除冲突豁免，评分却仍强制要求豁免例外条款，排除这一合格路线。

contracts/commercial-vendor-customer/master-services-agreement-first-draft/scenario-04

C1 答案或评分逻辑错误

公开指令要求内部编号放在页眉页脚，评分转换却丢弃这些位置，无法区分编号正确与缺失的正常文档。

intellectual-property/draft-international-trade-commission-section-337-complaint

C1 答案或评分逻辑错误

转让在投诉前已经完成，评分仍强制至少评为高风险，排除说明历史缺口后给出中等或已解决判断的有据分析。

Doubtful: 9 道题

firm-knowledge/tasks/097

C2 存在题面未说明的评分要求

C-006可能排除有证据效力的1010公开红线版本；目前尚未确认只提交该版本能否获认可。

emerging-companies-venture-capital/draft-stockholder-consent-resolutions

C1 答案或评分逻辑错误

评分要求指出的董事会同意障碍，似已被本版章程避免；正确说明“现版没有该风险”是否能通过尚不明确。

---

litigation-dispute-resolution/identify-issues-in-counterparty-interrogatories

C1 答案或评分逻辑错误

评分混用了专家报告提交期限与取证完成期限；正确区分日期的答复能否获认可，尚未确认。

---

environmental-esg/identify-issues-in-counterparty-contribution-claim

C2 存在题面未说明的评分要求

评分包含公开附件没有的两笔既有付款，本次也出现失败；尚未确认具体金额是否构成必要扣分条件。

---

litigation-dispute-resolution/draft-litigation-discovery-responses

C2 存在题面未说明的评分要求

评分预设未提供的运输日志有利于抗辩；是否接受诚实保留未知的条件式策略，尚不明确。

---

contracts/corporate-ma/merger-agreement-term-negotiation

C2 存在题面未说明的评分要求

邮件对5%终止费特批与4%限制的关系不清，评分又要求特定救济组合；不能排除该组合确有特批。

---

employment-labor/draft-eeoc-position-statement

C1 答案或评分逻辑错误

评分要求用替补员工同属少数族裔来削弱种族歧视推论，相关判例给出反向线索；是否允许审慎讨论尚不明确。

---

bankruptcy-restructuring/compare-competing-bid-terms-against-stalking-horse-asset-purchase-agreement

C3 关键错误漏判

评分可能放过仍保留Phase II条件、只补保证金就获准入的竞标方案；现有正常作答未触发这一疑点。

---

contracts/ip-licensing/license-agreement-first-turn-redline/scenario-01

C2 存在题面未说明的评分要求

评分列出内部60日退出条件，与对外函保密要求有冲突线索；是否接受兼顾保密的表达尚未确认。

评测影响与后续更新

这些问题会让模型因正确读取材料、采用合理起草方案而失去认可。纠正单个错误项未必改变某次总分，每项问题对实际评分的影响，取决于相应的触发条件。

后续如有作者补充材料、反证或修复，我们将按具体题目更新结论。

联系我们

benchmarks-last-exam.com Benchmarks-Last-Exam@outlook.com

HarmonyOS Sans SC · Copyright 2021 Huawei Device Co., Ltd.