GDP.pdf Review Report

作者：Benchmarks' Last Exam Lab 2026-10-15

GDP.pdf 用来自十个专业领域的真实 PDF 文档考查模型的阅读和推理能力。模型需要从文档中找到相关证据，用它回答金融、工程或研究等实际工作中的问题。

摘要

我们审查了公开的 100 道题，其中 **40 道 Accepted**、**13 道 Doubtful**、**47 道 Refused**。

已确认的问题主要集中在参考答案与原文不符，以及评分添加了题面没有说明的要求。有些条款把"不要答错"扩大成"连准确的背景和排除说明也不能出现"，使本来正确的答案失分。

审查结论	题数	判断依据
Accepted	40	本轮已审范围内，未发现实质缺陷或有具体证据支持的未决疑点。
Doubtful	13	存在具体疑点，但正确解释、触发条件或评分影响仍待核实。
Refused	47	至少一条影响评测有效性的缺陷已经确认，并经独立反证复核。

审查范围与版本

本报告覆盖开源副本中的 100 道题，固定数据版本：7f76578

评分实现使用 surge-ai/gdp-pdf 的固定版本：7a72a51

报告更新于 2026 年 11 月 4 日。

问题分布

问题类型	Refused	Doubtful	合计
C1 答案或评分逻辑错误	21	4	25
C2 存在题面未说明的评分要求	25	9	34
C3 关键错误漏判	1	0	1
C4 评分可绕过 / 答案泄露	0	0	0

问题类型	Refused	Doubtful	合计
C5 其他问题	0	0	0
合计	**47**	**13**	**60**

每题按主要原因计入一类。

代表性案例

下降100个基点，参考答案却写成1,000

C1 答案或评分逻辑错误

题目：gdp-060

题目要求分析Airbnb财报，并用基点表示成本、行政费用占收入比例的年度变化。1个基点等于0.01个百分点。

附件第71页的表格给出：成本占收入比例从2023年的17%降到2024年的16%。按这两个表列比例计算，下降1个百分点，也就是100个基点。评分标准r9却要求回答下降1,000个基点。

1,000个基点相当于下降10个百分点，与表中的成本单项不符。这里确认的是参考标准的数值错误。

证据 2

- 原财报第71页成本表
- 题面与r9原文：数据第60行

核心分析任务没有公布给解题者

C2 存在题面未说明的评分要求

题目：gdp-009

公开题面只说可以修改题面、上下文或评分规则来解决某个问题，详细内容在上下文或评分部分。它没有提出任何需要完成的分析任务。附件是十张照片和一页地图，也没有补充提问。

评分标准r1却要求对十张照片逐一填写三类判断：是否有冰雹损伤、是否能归因于2017年5月8日的风暴，以及是否存在维护不当。

解题者看不到这些工作要求。即使正确识别照片内容，也无法从公开输入得知必须按这三个维度逐项交付，核心任务定义与评分要求脱节。

证据 2

- 照片与地图附件
- 公开题面及r1：数据第9行

准确解释排除项，也会被扣分

C2　存在题面未说明的评分要求

题目：gdp-064

题目要求找出城市和农村儿童贫血率都至少为70%的年龄组，再报告各组的城乡合并总率。符合条件的是6—11个月和12—23个月，总率分别为88.1%和73.7%。

正常答案先列出五个年龄组的筛选结果，把另外三组明确标为不符合，最后只报告两个合格组。评分标准r5却把不得误选其他组，写成回答中不得包含其他年龄组。

判分因此把这条准确的排除说明判为失败；其余七项全过，均分为0.875，整题未获全项通过。问题是额外限制了正常解释方式，而不是答案选错了年龄组。

证据 3

- 原论文第8页年龄组总率
- 第10页城乡分组图
- 题面与r5：数据第64行

营养缺乏的后果整列写错，评分条款仍无法区分

C3　关键错误漏判

题目：gdp-094

题目要求从附件提取孕妇需要的营养素及缺乏后果，整理成两列表格，供儿科护理同事使用。原文先讲孕期铁、叶酸、碘缺乏的后果，随后另起一句讲儿童期维生素A缺乏与失明。

相关评分条款检查表格、列标题、三个营养素名称和不额外推荐补充剂，却没有核对后果列。把三个孕期营养素的后果都误接为相邻儿童期条目的失明，仍满足这些明文条件，其余条款也不检查这处改动。

被漏掉的是题目明确要求的营养素与后果关系。

证据 2

- 原附件第10页营养说明框
- 题面及全部21条标准：数据第94行

逐题审查结果

Refused：47 道题

gdp-002

C2　存在题面未说明的评分要求

题面只要求指出收费信息所在页码，评分却要求完整复述四类收费。页码定位正确、每页也给了真实例子，仍可能不符合这项额外要求。

gdp-004

C2 存在题面未说明的评分要求

变压器名称和零件号60536都正确，只因补充引用另一页的120V图，也被评分当作错误。该额外的引证限制实际使唯一一项失败，均分为0.75。

gdp-006

C1 答案或评分逻辑错误

参考答案把PESI风险评分138归为Class IV，附件表格和原始研究均应归为Class V。按来源正确分类反而与评分目标冲突。

gdp-009

C2 存在题面未说明的评分要求

公开题面只有一段编辑备注，附件也没有问题指令；评分却要求对十张照片逐一分析冰雹损伤、指定风暴归因和维护不当。核心任务没有告诉解题者。

gdp-013

C2 存在题面未说明的评分要求

答案已把体能状态未知者单列为Unknown，没有混入体能梯度，评分仍因表中出现这类信息而扣掉唯一一分。题面没有禁止这种准确说明。

gdp-015

C1 答案或评分逻辑错误

评分固定要求保险人承担抗辩义务，却忽略附件中针对袭击、殴打相关诉讼及过失理论的免责条款。忠实按保单作条件分析，会与该固定结论冲突。

gdp-018

C1 答案或评分逻辑错误

评分把水浸泡和加速老化两种试验条件混在一起。原表水浸硬度变化为±10，老化为+15/−0；实际判分因此扣掉了已答对的水浸硬度项。

gdp-021

C2 存在题面未说明的评分要求

题面允许按日期和 / 或问题给供应商排名，评分却禁止只按日期排名，并固定最差供应商。原表支持的另一种日期排名被排除。

gdp-022

C1 答案或评分逻辑错误

评分把承载关系写反，要求认可抽油烟机能支撑墙体。原文说的是弱墙上不装壁柜、抽油烟机属于例外，不能推出相反的结构承载结论。

gdp-023

C1 答案或评分逻辑错误

题目问哪类项目增幅最大，参考答案接受Power & Regional；原图中Neighborhood & Community从2023年末到2025年末的增幅更大。

gdp-024

C2 存在题面未说明的评分要求

评分把在本机构工作8个月，等同于累计合格联邦服务只有8个月。题面没有排除历史服务，手册也允许跨机构累计12个月，固定的不合格理由缺少依据。

gdp-025

C1 答案或评分逻辑错误

评分要求指出图纸缺少屋顶坡度，但附件第2页已在两侧屋顶标明6:12。把已经提供的信息当成缺项，会误判正确的图纸审核。

gdp-026

C2 存在题面未说明的评分要求

报告未给出待登记交易如何进入估值模型，评分却固定要求预测估值保持不变。依据现有信息说明无法判断方向的正常回答，被排除在该项之外。

gdp-028

C2 存在题面未说明的评分要求

正常答案在准确的页码导航中引用了含Mr. Henley的真实标题，没有把他误列为执行副总裁；评分仍因名字出现而实际扣分。

gdp-031

C2 存在题面未说明的评分要求

评分把HFCS暗中限为两种同名成品，排除按配方中高果糖浆属性判断产线资格的解释。两种读法影响28种产品的排产约束，属于核心建模条件。

gdp-036

C2 存在题面未说明的评分要求

四项目标名单完整正确，只多写一行准确的图纸版本来源，也会违反评分的禁入条件。题面没有禁止这类来源说明；现有轨迹未实测该错误拒绝。

gdp-040

C1 答案或评分逻辑错误

陶瓷轴承外圈的参考零件号写成316L616，原厂主图和两个外圈图标均为361L616。正确抄录零件号会与三条评分要求冲突。

gdp-041

C2 存在题面未说明的评分要求

正常答案已区分生育前后的工龄，并给出正确医疗费用，仍因补充0—4年工龄档信息而实际失分。题面没有这项禁提要求。

gdp-042

C1 答案或评分逻辑错误

评分把18—20岁组的BMI增速下降，写成BMI没有增加或变化。原表显示BMI仍在增加，只是增加得更慢，混淆了水平变化和速度变化。

gdp-043

C2 存在题面未说明的评分要求

评分要求识别并移动或避让嵌入式天花扬声器，但公开材料没有提供这一特定设备。忠实按图辨识实际障碍物的回答，仍会缺少该项指定内容。

gdp-045

C2 存在题面未说明的评分要求

题目中的type可以按PDF明确标出的TYPE列理解，得到13扇一层双悬窗；评分却暗中把它限定为MARK编号，并只接受0。

gdp-048

C2 存在题面未说明的评分要求

开放的初步电路规划被评分暗中限定为12—16个回路、至少12—14个配电位，排除了有理由支持的8个现用加2个备用位方案。

gdp-052

C2 存在题面未说明的评分要求

评分强制要求给出某条法律规定的起诉前至少30天需求函建议，附件没有提供该规则，题面又禁止使用附件外法律资料。

gdp-053

C2 存在题面未说明的评分要求

评分一律禁止反平方律计算，排除了利用论文50英尺远场数据、合理微调到15米的估计方法，即使数值符合要求。现有轨迹未实测这条方法的扣分。

gdp-056

C1 答案或评分逻辑错误

评分把Kokkonen设为唯一共同引用，排除了同样被共同引用、且已准确标明的下级判决103 F.4th 294。正常回答因此实际失分并未获全项通过。

gdp-057

C1 答案或评分逻辑错误

评分把四个中频模型的关联疾病仅归为PTSD，遗漏论文表5中的其他四种疾病，进而强制要求并非所有相关疾病共有的两种药物。

gdp-060

C1 答案或评分逻辑错误

财报表列成本占收入比例从17%降至16%，应下降100个基点；参考答案却要求1,000个基点，把单项变化写大了十倍。

gdp-061

C2 存在题面未说明的评分要求

正常答案的字段清单完整，只加了便于阅读的分组标题，便触发未公开的标题禁令而实际扣分。内容正确与否被额外排版条件混在一起。

gdp-062

C1 答案或评分逻辑错误

评分把Chloe没有在保单列名，当成全部基本医疗保障均不适用的充分理由，忽略附件中的法定一致条款及相关同住亲属保障规则。

gdp-064

C2 存在题面未说明的评分要求

答案正确选出两个年龄组，也准确解释了另外三组为什么不入选。评分把不得误选扩成不得提及，实际扣掉唯一一项，均分为0.875。

gdp-067

C1 答案或评分逻辑错误

参考答案把空调A-coil更换年份写成2020，原PDF第4页明确记为2022年9月。直接抽取原文日期会与评分真值冲突。

gdp-070

C1 答案或评分逻辑错误

参考答案接受2019年成交均价236,227.50美元，正确值为229,241.67美元。错误值来自把一笔2018年成交计入2019年。

gdp-071

C1 答案或评分逻辑错误

评分要求把堆土区低侧的防淤围栏判为不清楚或仅部分满足，原图C-3却已画出并用SF标明。正确识别已有围栏会偏离参考结论。

gdp-073

C1 答案或评分逻辑错误

评分把五次四碱基CTGC误并入三碱基TGC，要求CTGC为0、TGC为12；按原表下划线读取，应分别为5和7。

---

gdp-076

C2 存在题面未说明的评分要求

评分禁止提供男性相对死亡率的百分比增幅，排除了由原图支持、且明确标为近似值的正常比较。题面没有这项禁令，现有轨迹未触发该误拒路径。

---

gdp-079

C1 答案或评分逻辑错误

题目问急性ST段抬高型心肌梗死的心电图，评分却把几个导联压低固定为必答，排除有原始病例和队列依据的相应抬高表现。现有正常轨迹未实测这四项误拒。

---

gdp-080

C1 答案或评分逻辑错误

评分把满15年后才适用的每月1.75天假期费率，倒用于此前整整两年，要求42天；按题面时间轴分段计算通常约为36天。

---

gdp-082

C1 答案或评分逻辑错误

评分要求回答文中没有明确的方法优缺点，但附件第一页摘要已经给出深度学习方法的优势，以及双层解析拟合的优点和局限。

---

gdp-087

C2 存在题面未说明的评分要求

题面要求提取全部表格中的弱推荐，评分却暗限为表2A—2D，并禁止其他表里同样标为Category II的三条内容。按题面完整抽取反而不符合评分。

---

gdp-089

C2 存在题面未说明的评分要求

第5案没有给出首次实际现金价值付款日和维修费用发生日，无法计算条款的180天期限；评分仍固定判定无权追回保留款，并误用另两个日期替代。

---

gdp-091

C1 答案或评分逻辑错误

比较第二季度盈利能力时，参考答案用了行业上半年亏损机构占比6.30%。行业同季度应为6.59%，与社区银行6.75%的差距因此被错误放大。

---

gdp-093

C2 存在题面未说明的评分要求

表16列的是改进机会，没有给出现有标准的强度排名。评分却要求认定Mfg Pyramid目前最强，把机会覆盖误当成已有能力。

---

gdp-094

C3 关键错误漏判

题面要求总结孕期营养缺乏的后果，评分却只检查表格、名称等条件。把整列孕期后果误写成相邻儿童期条目的失明，仍满足全部相关明文条件；未实际运行该错误答案判分。

---

gdp-097

C2 存在题面未说明的评分要求

题面模板要求每小时产量预测，正常答案给出了可独立复算的1,595.8和1,665.5箱 / 小时；评分却额外要求预测周总箱数，实际拒绝了这两项成果。

---

gdp-098

C1 答案或评分逻辑错误

评分把涉及公共资金的记录一律不能适用律师通信豁免作为答案，遗漏原文对非特权内容的限定，也忽略保密法律账单仍可受保护的情形。

---

gdp-099

C2 存在题面未说明的评分要求

评分把工艺参数回答限制为一般公式，额外禁止附件允许的定性速度和进给建议。正常答案没有具体速率数值，仍因此实际失分。

---

gdp-100

C2 存在题面未说明的评分要求

正常答案使用了论文自身的R / I / E机制名称并作解释，评分却禁止这些词，实际扣分。禁用有来源的术语不是公开任务要求。

Doubtful: 13 道题

gdp-019

C2 存在题面未说明的评分要求

正确计算后的19.275按两位小数为19.28，参考为19.27；94.53%也可能超出评分的字面上界94.5%。现有轨迹未正确提交这些数值，接受范围仍待确认。

---

gdp-029

C2 存在题面未说明的评分要求

评分可能禁止在人数趋势段补充其他图中的金额数据，即使旁证标注清楚、人数分析正确。公开任务是否要求这种严格分隔，还没有确定。

gdp-030

C2 存在题面未说明的评分要求

仪器误差为±50 ppm，部分CO_2读数只略高于50 ppm，评分却固定要求据此判断六小时pH控制是否达标。误差范围内存在不同真实情况，结论是否可唯一确定仍有疑问。

gdp-033

C2 存在题面未说明的评分要求

题面没有交代既往牙周治疗史，评分却固定要求一次服务自付70%。有条件说明的答案已获分，削弱了误杀担忧，但该比例是否依赖缺失事实仍未厘清。

gdp-047

C2 存在题面未说明的评分要求

题面问两个数字的差，没有规定取绝对值还是按哪个顺序相减。17 / 22与−17 / −22均有解释空间，评分却固定要求负值并禁止正值。

gdp-058

C2 存在题面未说明的评分要求

负责人换任后的薪酬同比可以按职位或按个人计算。正常答案明确采用年末负责人并补充同人比较，仍因未用评分指定的进入 / 退出口径失分，其唯一性尚未确立。

gdp-063

C2 存在题面未说明的评分要求

目标年份结果已正确报告，额外两行明确标为基线背景，仍被唯一一项扣分。题面没有明示禁止准确背景，但是否超出要求范围尚有争议。

gdp-066

C1 答案或评分逻辑错误

评分把邮轮运营天数对收入增长的5%贡献，写成运营天数本身增长5%。现有回答另有足够的正确运营指标，尚未隔离出这处混淆对得分的影响。

gdp-074

C2 存在题面未说明的评分要求

题面要求扩大的蒸汽重整器边界，评分对401、428、430等流股的输入归属却采用固定口径。现有回答还包含其他错漏，尚不能把失分归因于边界歧义。

gdp-083

C1 答案或评分逻辑错误

将房车作为长期住所是否影响无保险驾驶人人身伤害保障，保单解释尚未唯一确定。正常答案的条件分析在三项失分，但另一项接受了保留意见。

---

gdp-085

C1 答案或评分逻辑错误

论文中的总体细胞扩增稳定，与EdU增殖检测增加可能同时成立，评分未清楚区分。两次本地失分加强了疑点，但仍可能存在兼容该项的完整正确答案。

---

gdp-088

C1 答案或评分逻辑错误

评分将competitiveness作为负面解释示例，可能误罚报告支持的特定领域竞争力目标；结合完整语境仍有相容解释，现有轨迹也未触发这一表达。

---

gdp-092

C2 存在题面未说明的评分要求

公开问题要求找出三种有前景的缓解方法，评分却固定从五项名单中选择。名单外的合理第三项是否会被误拒，现有运行还没有检验。

Accepted：40 道题

gdp-001
gdp-003
gdp-005
gdp-007
gdp-008
gdp-010
gdp-011
gdp-012
gdp-014
gdp-016
gdp-017
gdp-020
gdp-027
gdp-032
gdp-034
gdp-035
gdp-037
gdp-038
gdp-039
gdp-044
gdp-046
gdp-049
gdp-050
gdp-051
gdp-054
gdp-055

gdp-059

gdp-065

gdp-068

gdp-069

gdp-072

gdp-075

gdp-077

gdp-078

gdp-081

gdp-084

gdp-086

gdp-090

gdp-095

gdp-096

评测影响与后续更新

文档问答的评分标准需要忠实于原文，也需要接受题面允许的合理回答。错误的数值、日期或对象会把正确提取与得分方向倒置；隐藏的禁词和排他要求，则会让准确解释变成扣分理由。

相关领域判断仍需专家进一步核实。每项问题对实际评分的影响，取决于相应的触发条件。后续如有可核查的反证、作者澄清或修订，我们会更新相应结论。

联系我们

benchmarks-last-exam.com

Benchmarks-Last-Exam@outlook.com

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