# AA-Omniscience v1.1 Review Report

作者：Benchmarks' Last Exam Lab 2026-11-02

AA-Omniscience Public v1.1 是Artificial Analysis发布的知识评测公开子集，覆盖科学工程、软件、金融、健康、人文社科和法律，要求模型闭卷回答短答案问题，也允许不知道时弃答。

## 摘要

**我们从公开版本的 300 道题中分层抽查了 100 道，其中 98 道 Accepted、1 道 Doubtful、1 道 Refused。**

确认的问题出现在一道会计题：指定资料中有两个对象符合题目描述，题面未作区分，答案却只接受其中一个。另有一道心脏发育题存在文献解释的具体疑点。

| 定档 | 数量 | 含义 |
|---|---|---|
| Accepted | 98 | 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 | 1 | 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 | 1 | 我们已确认至少一项影响评测有效性的问题。 |

## 审查范围与版本

使用AA-Omniscience Public v1.1, 1f9cbcec4720，按领域与子主题分层抽样，科学工程与数学、软件各20题，金融、健康、人文社科、法律各15题。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 | 0 | 1 | 1 |
| C2 存在题面未说明的评分要求 | 1 | 0 | 1 |
| **合计** | **1** | **1** | **2** |

## 代表性案例

### 两个符合描述的利润小计

C2　存在题面未说明的评分要求

题号 122 **Refused**

题目问IASB在2018年9月项目概览中，与EBIT一起提出的新利润表小计叫什么。参考答案为Business profit。

指定资料第10页却同时列出Business profit、Profit before investing, financing and income tax, 以及EBIT。前两个都符合题目描述，而且不是同一个小计的长短名称：第12页例表分别给出3,000、3,500和3,900，差别涉及一体化联营与合营企业利润是否计入。

题面没有给出排除第二个对象的条件，评分却只认一个。

证据 1

- IASB project overview, slides 10 and 12

---

### 心脏分隔的起始阶段

C1　答案或评分逻辑错误

题号 166 **Doubtful**

题目问人类胚胎心脏远端流出道何时开始分隔、何时形成独立的主动脉和肺动脉干。参考为Carnegie第16、18阶段；模型回答第14、16阶段，被判错误。

疑点在第一问：2003年论文图注把已分隔标本标为第14阶段，2012、2018年研究则支持第16阶段出现实际腔道分隔。

证据 3

- Webb et al., 2003
- Sizarov et al., 2012
- Yang et al., 2018

---

### 格律答案的部分给分

题号 220 **Accepted**

题目要求给出波斯诗格律代码4.5.11的名称，参考为khafif。模型答出更完整的Khafīf-e musaddas-e makhbūn-e maḥḍhūf，并附音步和节律符号，原判为部分回答。

完整名称、音步和节律符号均正确。但题目只问名称，作答指令要求只给答案、不解释；附加音步和符号超出命名所需内容，部分给分有合理解释，复核后不再作为单题疑点。

证据 2

- The Individuality of the Persian Khafīf Metre
- AA作答与判分规则

### 细胞体遮光的依据

题号 5 **Accepted**

题目问两种没有可检测眼点的甲藻，靠什么结构在趋光时遮光。参考为Cell body，模型回答Chloroplasts，被判错误。

教材直接点名这两个物种，说明细胞体承担遮光作用。此前质体遮光的依据来自对其他甲藻的一般性讨论，不能据此证明本题回答“叶绿体”应该得分，复核后撤销疑点。

证据 2

- Algae: Anatomy, Biochemistry, and Biotechnology，第97页
- Jékely, 2009

## 逐题审查结果

### Refused: 1 道题

122

C2 存在题面未说明的评分要求

两个不同利润小计均符合题意，题面缺少区分条件，固定答案只接受其中一个。

### Doubtful: 1 道题

166

C1 答案或评分逻辑错误

起始分隔阶段有文献口径差异；较新研究支持参考答案，尚未确认错误。

## 评测影响与后续更新

本次样本中，大多数题目未发现实质缺陷。

后续如有作者补充来源、判分解释或修复，我们将据此更新对应题目的结论。

### 联系我们

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

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