

湖北省第二届“数字工匠”职工技能大赛
数据标注员赛项技术文件

2025 年 7 月

目 录

一、 命题原则	1
二、 比赛形式	1
三、 比赛内容	1
(一) 比赛时间	1
(二) 赛题分数	1
(三) 比赛题型	2
四、 评分规则	2
五、 必备工具	3
六、 技术平台	6
五、 样题	8

一、命题原则

数据标注员是人工智能发展的基石，高质量的标注数据直接决定 AI 模型的性能上限。依据国家职业技能标准，本次竞赛以“精准标注·AI 之基”为主题，聚焦数据标注员的实战能力与职业素养，旨在通过多维度、多模态的标注任务，选拔具备高效性、精准性及规范意识的数据标注技术人才。

二、比赛形式

本赛项为个人赛，分为数据标注理论赛（理论考试）和数据标注技能挑战赛（实际操作考试）两部分组成。比赛采取闭卷机考方式进行。赛事分为：理论考试、数据标注技能挑战赛（文本内容标注场景）、数据标注技能挑战赛（语音标注场景+图像识别场景）。

三、比赛内容

（一）比赛时间

上午场：数据标注理论赛+数据标注技能挑战赛（文本内容标注场景）；理论考试时长为 1 小时，数据标注技能挑战赛（文本内容标注场景）时长为 2 小时。

下午场：数据标注技能挑战赛（语音标注场景+图像识别场景），时长为 4 小时。

（二）赛题分数

本次竞赛设计理论实操分值占比为 3:7，竞赛总分 1000 分。其中，数据标注理论赛为 300 分，数据标注技能挑战赛 700 分（文本内容标注场景分数为 200 分，语音标注场景分数为 250 分，图像识别场景分数为 250 分）。

（三）比赛题型

数据标注理论赛：题型包含单选、多选题及判断题，随机抽取 100 道题。其中，设定单选题 50 题，多选题 30 题，判断题 20 题。主要考察参赛选手对于数据标注行业合规政策、隐私保护技术规范等核心内容的掌握情况，同时涵盖数据标注场景下的安全风险识别（如敏感信息泄露）、合规标注流程设计及脱敏技术应用能力。

数据标注技能挑战赛：题型设计为技能实操题。提供选手特定场景下的数据集和数据标注要求，选手通过相关技术手段完成对应场景的任务要求，并按任务要求规定的结果格式构建最终文件作为答案提交至平台验证。主要考察参赛选手对于实际业务场景下图像、文本、视频、语音等多模态数据的标注能力。

四、评分规则

数据标注理论赛：采用固定积分模式，参赛选手需根据题目描述，选择合适选项进行作答。作答过程中计时不中断，

选手需逐题点击提交。若在规定时间内（开赛1小时内）未完成全部题目作答，则平台自动判定交卷并计分。

数据标注技能挑战赛：采用自动化动态评分模式，参赛选手需根据提供的场景及标注要求进行数据标注操作，通过标注结果的达成率进行评分，选手标注结果的达成率越高，则对应任务的分值就越高，每个任务具有固定分值。

总成绩=数据标注理论赛得分+数据标注技能挑战赛得分，按照总分由高到低进行排序，最终成绩分数相同，则用时短者排名靠前。

五、必备工具

（一）比赛前需各参赛选手自行准备相关必要工具和脚本，包括但不限于以下相关工具：

- **yolov11** 模型训练环境，训练模型需兼容 **ultralytics/ultralytics:latest-cpu** 容器进行模型推理。

- 图像标注工具（**make-sense**、**Labelimg**）等。
- 语音标注工具（**Praat**）等。
- **python** 环境及常用库。

（二）推荐参赛电脑配置不低于如下配置：

- **CPU：**i7 12 代以上及同等配置 CPU。
- **显卡：**NVIDIA RTX 3060（8GB 显存）及以上。
- **内存：**16GB 或更高。

- 硬盘：SSD 100GB 以上剩余空间。
- 耳机：有线耳机。

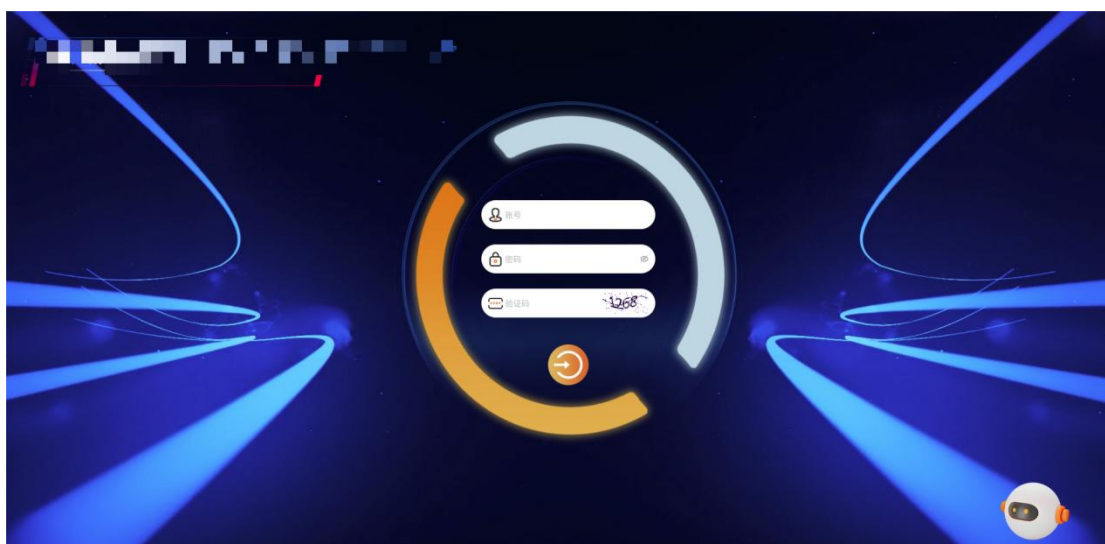
六、技术平台

（一）比赛平台

1.数据标注理论赛

理论赛答题流程：

①登录竞赛平台



②、选择 “数据标注理论” 模块



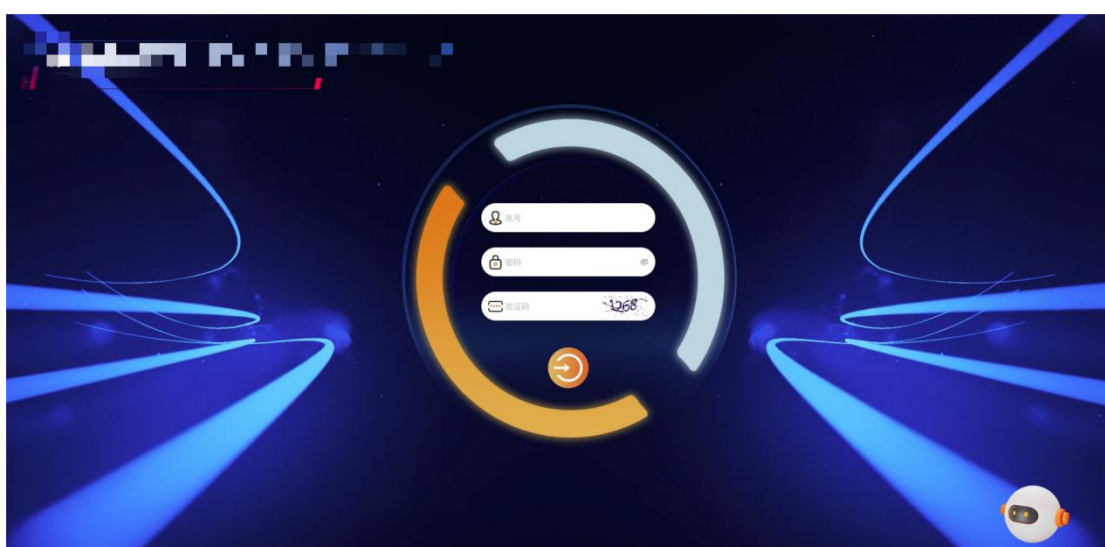
③、系统随机抽取 100 道题

考生可以自由选择题目作答，可以跳过暂时不答，每道题作答后需要点击“提交答案”。已作答的题目不能进行二次作答。

2.数据标注技能挑战赛

技能挑战赛答题流程：

①登录竞赛平台



②选择“数据标注技能挑战赛”模块



③题目作答

根据页面提示，点击访问环境 IP 地址或者下载相关附件。考生依据考题描述，点击“提交答案”按钮提交相关结果文件进行校验。系统会自动验证选手提交的内容并显示选手当前任务获取的分数。请留意每道题目的答题提交次数限制，超过限制次数后，不可提交答案！

3.观摩推送页面



七、样题

(一) 数据标注理论赛样题

1. 单选题

根据《个人信息保护法》，在标注涉及个人生物识别信息（如人脸图像、语音）的数据时，以下哪种做法符合合规要求？

- A. 直接标注原始数据，无需额外处理
- B. 删除人脸区域，并对语音进行变声处理

- C. 仅标注数据，不存储原始文件
- D. 删除用户 ID 但保留年龄、性别等属性

答案：B

解析：

生物识别信息属于敏感个人信息，需严格脱敏（如删除人脸、变声处理）。

C 选项不存储原始文件不符合数据留存要求，D 选项保留年龄性别仍可能泄露隐私。

2.多选题

以下哪些指标可用于量化数据标注质量？（多选）

- A. 图像标注交并比（IoU）
- B. 文本实体识别的 F1 值
- C. 分类任务准确率（Accuracy）
- D. 标注者间一致性（Cohen's Kappa 系数）

答案：A、B、D

解析：

C 选项“准确率”在类别不平衡时易失真（如 99%负样本），需配合精确率/召回率使用。

IoU（图像）、F1 值（文本）、Cohen's Kappa（多人标注一致性）均为行业通用指标。

3.判断题

数据标注完成后可直接用于 AI 模型训练,无需进行数据清洗和数据集划分。

答案：错误

解析：

标注后必须进行数据清洗（剔除噪声数据）、数据集划分（训练集/验证集/测试集），否则易导致模型过拟合或评估失真。

（二）数据标注技能挑战赛样题

1.样题任务场景：自动驾驶图像标注

【任务描述】

数据：城市道路图像（包含车辆、行人、交通标志、红绿灯等）。

【标注要求】

目标检测（Bounding Box）：标注车辆（car）、行人（pedestrian）、交通灯（traffic_light）。

语义分割（Polygon）：精确勾画道路区域（road）与非道路区域（如绿化带）。

【自动化评分】

目标检测：计算与标准答案的 IoU（交并比）， $\text{IoU} \geq 0.9$

得满分，按比例扣分。

语义分割：使用 Dice 系数评估分割区域重合度，阈值设为 0.85。

工具：COCO 评估工具（PyCocoTools）自动输出 mAP（平均精度）。

2.样题任务场景：医疗文本实体标注

【任务描述】

数据：电子病历文本（如“患者主诉持续性胸痛，ECG 显示 ST 段抬高”）。

【标注要求】

实体识别（NER）：标注疾病（disease）、症状（symptom）、检查项目（test）。

关系抽取：标注“症状-疾病”关联（如“胸痛→心肌梗死”）。

【自动化评分】

实体识别：基于字符级 F1 值
（`sklearn.metrics.classification_report`）。

关系抽取：精确匹配头尾实体及关系类型，正确率 $\geq 95\%$ 得满分。

工具：SpaCy 或 Flair 预训练模型自动比对标注结果。

3.样题任务场景：跨模态关联标注（高阶）

【任务描述】

数据：电商直播视频（含画面、语音、弹幕文本）。

【标注要求】

图像标注：识别主播手中商品（如“口红#色号 A12”）。

文本标注：提取弹幕中的购买意向关键词（如“下单”“求链接”）。

语音标注：转写主播促销话术中的价格信息（如“原价 399，现价 199”）。

【自动化评分】

跨模态对齐：商品名称、价格、弹幕关键词是否逻辑一致（规则引擎校验）。

多维度权重：图像（40%）+文本（30%）+语音（30%）加权得分。

工具：自定义 Python 脚本校验时间戳对齐与内容关联性。