高质量数据集建设指南

面向客户沟通、项目交付和销售赋能的数据服务方法

这份资料用于沉淀市场增长工作中的方法、判断和执行标准。内容以实操经验为基础，强调可复用、可检查、可交接。

一、高质量数据集不是数量堆叠

数据集建设的核心是让数据能够支撑模型目标。数量、覆盖度、一致性、准确率、合规性和可追溯性共同决定数据价值。

- 先定义模型任务和使用场景，再决定数据类型和采集范围。
- 先定义质量标准和验收方式，再安排采集与标注。
- 先确认合规边界和授权方式，再进入规模化执行。

二、建设流程

阶段	关键动作
需求定义	明确模型任务、数据类型、场景范围、标注粒度、交付格式和验收标准。
样本设计	覆盖地域、光照、设备、姿态、噪声、人群、长尾场景和负样本。
采集执行	制定采集SOP、设备标准、人员培训、现场质检和异常记录。
标注规范	定义标签体系、边界规则、示例库、冲突处理和版本管理。
质量控制	抽检、复检、交叉验证、误差统计、问题回流和人员校准。
交付验收	提交数据、标注文档、质检报告、问题清单和版本记录。

三、客户沟通中常见问题

- 客户只说要数据，未定义模型任务：需要反问使用场景、指标目标和失败样本。
- 客户只看报价：需要解释质量、合规、采集难度、交付周期和返工风险。
- 客户需求频繁变化：需要版本化管理需求，并明确变更影响。
- 客户忽视长尾场景：需要用样本覆盖度说明为什么数据不能只取容易采集的部分。

四、质检指标

指标	说明
准确率	标注结果与标准答案一致的比例。
一致性	不同人员对同一规则的执行稳定性。

覆盖度	样本是否覆盖主要场景、边界场景和长尾问题。
完整性	字段、图片、音频、文本、元数据是否缺失。
可追溯	是否能追踪数据来源、处理过程、标注人员和质检记录。
合规性	是否满足授权、隐私、脱敏和安全要求。

五、交付材料建议

- 数据需求确认表
- 采集SOP
- 标注规范书
- 示例与反例库
- 质检报告
- 问题复盘表
- 版本变更记录