选择人工 AI 数据采集的五大原因

训练模型需要数据，可以是合成数据、人工采集的数据，也可以两者结合。了解为什么使用人工采集的数据（至少是所用的大部分数据）能让模型性能更强大、更可靠。

1 质量

质量更高的数据能够确保模型获得更好的性能，因为这样的数据通常更准确、更新鲜、更干净、更一致、结构更完善，并且包含丰富的背景信息。所有这些特性使模型更容易对语言和相关主题形成更细致的理解，从而使其达到更高的性能水平。

示例

Lionbridge 为一家在线学习解决方案提供商提供支持，对 300 多段机器转录的视频进行了质量审核。审核人员修正了 AI 转录的错误，并提供了准确度非常高的视频字幕。

2 多样性

如果提供商数据贡献者遍布全球且高度多元化，其提供的数据就能更全面地覆盖现实场景，并反映多元化、无偏见的观点。

示例

为了服务一家大型科技公司，Lionbridge 依靠 Aurora AI Studio™ 和我们的招募团队，收集了超过一百万条特定情绪的讲话录音。参与者使用多种语言和方言，涵盖了各种社会群体。

3 速度

如果选择了合适的服务提供商，速度会非常快。Lionbridge Aurora AI Studio 连接着遍布全球的五十万名贡献者。如果需要，我们还有一支强大的招募团队，可以从贡献者群体之外寻找更多人才。我们发布的任务会在几天甚至几小时内被领取并完成。

示例

为了服务一家智能手机制造商，Lionbridge 使用 Aurora AI Studio 平台收集了一个涵盖 8 种语言的庞大“真实生活”对话数据集，其中包含超过 20 万段对话，每段对话最多有 5 位参与者。我们的平台让我们得以在 4 周内交付所有对话数据。

4 价格

仅使用合成数据通常会导致模型性能不佳，因此需要额外的验证程序或获取更多人工采集的数据。一开始就选择人工采集的数据可以降低成本。

示例

Lionbridge 为一家在线视频服务提供商提供支持，确保大量视频从多种语言翻译成英语。译员会标记所有粗俗或者带有冒犯性、偏见或仇恨的内容。这种细致入微的关注和审核使大型语言模型得以更快进入理想运行状态 — 更加优化且不含冒犯性内容，而无需进行额外的数据采集或验证。

5 开发流程

开发工作成本高昂。不要浪费开发人员的时间和人力成本。从一开始就选择可靠的人工采集数据。

示例

为了给一家 AI 开发商提供支持，我们指派了人工审核员，他们可以为大型语言模型提供多种语言所需的大量学习数据。他们从对提示的多条回复中选出最优回复，并根据多个因素对回复进行评分。

LIONBRIDGE

准备好探索 AI 数据采集服务了吗?立即联系 Lionbridge AI™！
访问 lionbridge.com 了解更多信息