数据标注行业校企合作路径探索

——以某多模态双工 QA 标注项目为例

王浩辰

天津外国语大学 天津 300204

【摘　要】：高质量训练数据是多模态大模型发展的关键瓶颈，而多模态双工问答等开放式标注任务对传统质量控制范式提出新挑战——评估具有强主观性、难以依赖量化指标。现有文献多聚焦于 MTurk 低成本众包或专家标注两种极化方案，对“校企合作—学生兼职”中间模式缺乏系统讨论。本文基于一项某为期 15 天、约 150 名学生兼职完成 1000 条多模态双工问答项目，提炼出适配开放式标注任务的“校企合作实践方法”，涵盖标注过程、人员管理、质量估计与改进等核心维度。研究表明，校企合作模式在伦理合规与质量上限上优于一般众包，在成本与培训弹性上优于纯专家标注，可作为面向高难度开放式标注任务的“第三条道路”。

【关键词】：数据标注；开放式众包；质量控制；多模态双工问答；校企合作；项目管理

DOI:10.12417/2982-3811.26.02.031

1 引言

近年来人工智能领域正经历从“模型驱动”向“数据驱动”的范式转变。GPT 系列从 GPT-1 的 4.6 GB 扩展到 GPT-4 约 40000 GB（13 万亿词元），Llama 3 训练数据已超 15 万亿词元，能够直接喂给模型训练的高质量数据集愈发成为产业链稀缺资源。我国 2024 年数据标注市场规模已突破 120 亿元，预计 2025 年达 200~300 亿元区间。进入多模态阶段，“双工问答”——用户与 AI 在同一视频流上进行连续、双向、可被打断、可被主动引导的对话——这类数据在公开互联网几乎不存在天然语料，必须通过人工“演绎—撰写—对齐—配音”等多步骤创作形成，本质上构成多模态对话模型能力上限的“数据瓶颈”。

开放式标注与分类型标注存在根本差异。分类型任务（实体识别、目标框选、情感分类等）通常存在唯一标准，可用标注一致性（IAA）度量；而开放式标注“答案空间巨大、并无唯一解”，两位标注员或会独立产出可能在表层文字上差异显著但语义上同等优秀的答案。而多模态双工问答则面临更多挑战——任务无单一正解，IAA 等指标失效；质量评估涉及视觉对齐、对话连贯、视角一致、交互密度等多维主观判断；验收要求“重做直至接受”导致质检成本陡增。

现有文献主要关注的项目方案有两类：

（1）MTurk 众包模式。此模式工人质量波动大、跨文化背景不可控、伦理风险突出；

（2）专家标注模式。该模式在质量上限上有明显优势，但成本高昂、规模受限、培训周期漫长。

本研究备考真实项目，探索“校企合作—学生兼职”的项目运营模式，利用高校学生的兼职意愿与基本素养组建中等规模团队，填补这一在学术文献中几乎从未被系统讨论的新路径。

2 文献综述

Roit 等（2020）针对问答式语义角色标注（QA-SRL）提出“受控众包”四要素，即筛选、培训、反馈循环、限定发布任务，从 30 名参与者中选定 11 名质检员（约 36%），使首次通过率提升约 25%。

Klie 等（2024）对 591 组标注数据的实证分析提出“标注过程、标注员管理、质量估计、质量改进、验收”的五维分类法，发现约 30%的数据质量管理不合格，对 IAA 的使用不够规范。但上述研究的样本以分类型任务为主，其框架主要面向“识别分类型”任务，并不完全适用于开放式标注项目。

Chai 等（2024）提出“任务—标注员—答案—系统”四要素质量模型，明确开放式众包与分类型任务的根本差异，但论述偏理论缺组织实践案例。

Braun（2024）等通过专家访谈+公司用户研究揭示“研究—实践鸿沟”，为本文实证立场提供启发。

Ahmadzadeh 等（2025）基于太阳科学图像标注实践，提出五条成功标准：C1 产出数量与质量、C2 形态与格式、C3 资源消耗、C4 目标服务、C5 项目文档化。

3 项目背景

本项目目标产出 1000 条高质量多模态双工问答数据，团队约 150 人，周期 15 天。每条数据时长 30 秒至 3 分钟，需覆盖至少 2 种能力标签（数量统计、OCR、世界知识、实时描述、细粒度动作理解、主动/被动提醒、记忆、打断、多轮对话、推理、声音互动、主动引导、主动纠错、手势交互、步骤指导等）。多模态双工问答的复杂度体现在三个维度：视觉模态实时性（“一分钟以下视频连续静默不超过 5 秒，一分钟以上不超过 15 秒”）；对话双向性与拟人感（要求体现共情与自然停顿、

严禁穿帮）；视角与人称的全程一致性。每条数据本质是小型对话剧本创作，单条标注实测耗时约 40 分钟，反映出开放式标注的“内在高成本”。

4 标注过程设计

本次项目的 15 天紧凑周期使得项目各环节必须同步并行。在这一约束下，我们采取了两种策略：

其一是前置投入比例的提升，即将培训与试点合并，把更多原本应分布在“试点轮次”的成本前置到培训阶段，以最大化保证正式生产的进度。传统的试点研究通常占据项目周期的 10%-20%，在本项目中，这一时长（约 1.5-3 天）若用于独立试点将严重挤压生产时间。我们采取的折中方案是“培训—试点合一”：在 3 天培训期内，要求每位通过初阶段培训的标注员完成 2~4 条小批量标注样本，这些样本既作为标注员的“实战练习”，也作为团队的“项目级试点”——领域顾问从这些样本中识别出需要在标注规范中明确的边界情境，以新增的 FAQ 条目形式更新规范文档。

这一设计的优势是显著的时间复用：同一批样本既服务于个体训练，又服务于项目级试点。其代价是失去了“独立试点”所能提供的“零基线偏差”——正式标注员在练习中产出的样本，本身已带有“刚学完培训的风格偏差”，其在样本难度分布上也未必能代表后续大规模数据。

其二是滚动修订机制，标注规范不在“试点结束”才修订，而是每天由质检员基于前一日质检反馈进行小幅迭代。这种“高频小步迭代”的设计虽然失去了 Klie 等所倡导的“完整试点”的统计优势，但在短周期约束下是更现实的选择。本项目的标注规范在 15 天周期内经历了多个版本的演化。初始版本主要基于项目启动前由项目方与领域顾问共同起草的需求文档，涵盖标注方式、流程、单条时长、交互密度、对话感、推理能力问题设计、视频切片选取规则、能力标签详细说明等核心条目。在培训—试点阶段，主要变化集中在“边界情境”的明确化，例如“何种情况下 AI 可以不输入用户文本而直接输入回答”“第一人称视角下主人公能否自言自语”等。在生产阶段中后期，主要变化集中在质检反馈高频出现的问题条目的细化。

5 质量控制

在开放式标注任务中，IAA 的概念不再适用于质检。IAA 适用于度量“对同一对象的判断是否一致”，不适用于度量“各自创作的产物是否同等优秀”。而在开放式标注中，两组围绕同一视频但主题迥异的优秀产出会被误判为“严重不一致”。

本项目采用“全盘质检”，原因有二：抽样质检在 1000 条规模下置信区间过宽（5%抽样仅 50 条样本，统计意义有限）；精品数据对单条质量极端敏感，任何错误数据都可能成为模型训练的“反样本”。建构五维主观指标体系——视觉对齐度（指代是否匹配画面、实时描述是否同步）、对话连贯性（上下文衔接、视角一致）、回答自然度（避免书面腔、共情力、合理打断）、信息密度（避免无意义寒暄、覆盖能力标签、推理链完整）、格式合规性（音色、时长、标签数量）——采用“任一维度不达标即退回”的短板原则，避免量化精确性幻觉同时保留可解释性。质检员—审核员二级分工将“客观合规判定”与“主观卓越判定”分流：审核员把“基础合规线”，质检员把“卓越质量线”，1:2 配比既保证深度判断时间又避免成本膨胀。

为了在开放式生成式任务上实施全盘质检，本项目建构了一套多维质量评估指标体系。该体系包含五个一级维度，每个维度下又细分若干二级评估点。

视觉对齐度：评估问答内容是否准确对应视频画面；包括“问题中的指代是否匹配画面元素”“回答中的实时描述是否与画面变化同步”“人物动作的描述是否与实际动作一致”等二级点。

对话连贯性：评估多轮交互的逻辑流畅性；包括“上下文是否自然衔接”“话题转换是否合理”“第一/第三人称视角是否全程一致”等。

回答自然度：评估对话是否符合人类自然交流习惯；包括“是否避免书面腔”“是否体现共情力”“是否合理使用打断与主动引导”等。

信息密度：评估回答的实质内容含量；包括“是否避免无意义寒暄”“是否覆盖所要求的能力标签”“推理类问答的推理链是否完整”等。

格式合规性：评估交付物是否符合工程规范；包括“语音轨音色是否符合要求”“时长是否在规定区间”“单条是否覆盖至少 2 种能力标签”等。

这套指标体系的核心特征是多维度、可言说但不可数值聚合——每个维度都有清晰的评估点供审核员判断，但其结果不被强行折算为单一数值或加权平均，而是采用“任一维度不达标即退回”的“短板原则”。这一设计避免了将主观评估强行量化所带来的精确性幻觉，同时保留了评估结果的可解释性。它在功能上替代了 IAA 在传统任务中所扮演的“质量基线”角色，但在概念上是完全不同的工具。

6 质量改进

本项目的质量改进机制核心是“重做直至接受”（Redo Until Accepted），具体实施包括以下要点：

退回理由分类：审核员/质检员退回数据时，必须从五个一级维度的二级评估点中明确选择具体退回理由，并简要说明问题所在。这一规定避免了“凭感觉退回”导致标注员无所适从。

修改时限：被退回的数据需在 24 小时内提交修改后的版

本，以维持项目整体进度的稳定性。

最大返工次数：同一条数据的最大返工次数被设定为 3 次；若 3 次返工后仍不达标，则该条数据将被记录，标注员的当条工时按一半结算，多次出现这种现象将剥夺其标注资格。这一上限设计的目的是防止个别困难任务陷入无限返工循环，既保护项目进度，也保护标注员积极性。

反馈的有效性在很大程度上取决于其颗粒度。本项目对审核员/质检员的反馈格式做出统一要求：必须对事不对人（评论数据本身，不评论标注员个人）、具体到轮次（说明哪一轮对话的哪个维度有问题）、给出修改方向（不只是“这里不好”，而是“这里应当改为类似 XXX 的形式”）。

实践中我们观察到，标注员对反馈的吸收会从开始的“被动”逐渐转变为“主动内化”。初期标注员往往机械地执行反馈（“被告知改 A 就改 A”）；中期开始理解反馈背后的标准（“意识到 A 的问题源于对标准的误解”）；到后期，对标准的内化理解会让标注员的返工率大幅下降。

7 研究局限

由于本文的方法论建构基于单一项目，虽然该项目在多模态双工问答这一任务上具有典型性，但其结论向其他生成式任务（如代码生成、长文写作、对话评估）的推广需要进一步验证。本项目涉及的双工问答属于“创作型+实时多模态”的特殊任务，本框架在更纯文本类的开放式任务（如长摘要、复杂改写）上的适用性需要进一步检验。

另外，本项目的标注员主要来自中国大陆某校园招聘渠道，文化背景相对集中，跨文化推广（如东南亚高校、欧美高校）需谨慎。

8 结论与展望

在系统对比了 MTurk 众包、校企合作-学生兼职、专家标注三种模式后，本研究认为校企合作模式在标注员稳定性、文化背景可控性、反馈循环有效性、质量控制等方面优于 MTurk 众包；在成本结构、规模弹性、培训弹性上优于纯专家标注，可作为面向高难度生成式标注任务的“第三条道路”。

基于本研究的发现，未来研究可以重点关注人工智能辅助标注与开放式标注的合作作业方向。目前主流的人工智能辅助标注方案多面向分类型与分割型任务，开放式标注的人工智能辅助路径也具有很高的探索价值。例如，“人工智能预生成+人类精修”的人智协同标注；多维主观指标体系是否可被部分自动化；视觉对齐度是否可通过视觉问答模型进行预筛；对话连贯性是否可通过语言模型进行评估等等，对这些路径的开发对于降低全盘质检的人力成本极具价值。

另一方面，学生标注员能力认证体系也存在建设与优化价值。本项目中涌现出的优秀标注员们证明了数据标注也可以作为一种实用技能教授给学生，而数据标注产业的人力缺口在未来会长期存在。因此，学界可以设计一套标注员技能的“段位认证”体系，为产业的规范化做出贡献。

参考文献：

[1] 燕江依等.信息通信技术与政策,2025,51(8):26-34.

[2] Chai L,et al.Quality Control in Open-Ended Crowdsourcing:A Survey.arXiv:2412.03991,2024.

[3] Ahmadzadeh A,et al.A Guide for Manual Annotation of Scientific Imagery.arXiv:2508.14801,2025.

[4] Roit P,et al.Controlled Crowdsourcing for High-Quality QA-SRL Annotation.ACL 2020.

[5] Klie J C,et al.Analyzing Dataset Annotation Quality Management in the Wild.CL,2024,50(3).

[6] Braun T,et al.Understanding the Process of Data Labeling in Cybersecurity.SAC 2024.