昆明理工大学

实习报告

学　　院：信息工程与自动化学院

专　　业：数据科学与大数据技术

年　　级：2021级

学生姓名：蒋　学　琛

指导老师：李　亚

日　　期：2025年3月20日

教务处制

目 录

一、 引言

1.1 实习背景与任务概述

人工智能（AI）特别是大型语言模型（LLMs）的飞速发展，对高质量、精对齐的训练数据提出了前所未有的需求。字节跳动的“豆包”大模型即是此类先进模型的代表。本次实习于文远知行科技有限公司进行，核心任务是作为大模型实习生，参与为“豆包”模型提供标准化、可信的数据支撑项目，通过一个严谨的多轮数据处理与对齐流程，确保数据质量。这对于理解 AI 产业中数据的核心作用及掌握前沿数据处理技术具有重要意义。文远知行在 AI 及数据处理方面的深厚积累为实习提供了坚实基础，字节跳动作为合作方，其“豆包”模型对数据的高标准严要求则为实习设定了明确目标。

1.2 实习岗位与主要职责

实习岗位为 AI 数据平台部大模型实习生，主要职责包括：理解“豆包”模型对数据的具体需求（特别是对齐标准）；参与并执行三轮核心数据处理流程（初步清洗、两轮核心对齐、最终质检与标准化）；协助数据采集与预处理；执行数据标注与审核；参与数据质量评估与反馈；进行跨团队沟通并撰写相关技术文档。

二、 实习内容与过程

2.1 大模型数据需求理解与分析

实习初期，通过研读规范文档、参与需求会议及与资深工程师交流，深入理解了“豆包”模型对数据类型（SFT、RLHF、预训练语料等）、格式（JSON/JSONL）及核心“对齐”目标（事实性、逻辑性、指令遵循、风格展现、无害性、价值观符合等）的要求，明确了评价指标与验收标准，为后续工作奠定了基础。

2.2 三轮核心数据处理流程详解

为确保数据质量，我们严格遵循一个三轮核心数据处理流程：

2.2.1 第一轮处理：原始数据清洗与粗粒度对齐

核心活动：整合多源原始数据，基于规则（关键词、长度、纯净度）进行初步筛选；利用 Python 脚本（Pandas, Re）自动化清洗（去标签、特殊符号，规

范化，去重）；根据初步对齐指南进行格式统一和基础内容筛选（如安全合规）。**产出：**初步净化和粗对齐的数据池。

2.2.2 第二轮处理：深度对齐与精细化标注/修正

核心活动：针对 SFT 阶段，构建“指令-输出”对；为 RLHF 阶段准备偏好数据（排序或打分）；对数据进行内容改写与优化（提升清晰度、准确性、逻辑性）；进行多轮人工审核与校验，确保标注准确性与一致性；与字节团队就样本进行迭代反馈。**产出：**高质量、高对齐度的数据集。

2.2.3 第三轮处理：最终质检、标准化封装与可信度增强

核心活动：自动化与人工结合的全面质检（格式、字段、逻辑）；严格按最终格式标准化封装；采取可信度增强措施（如事实核查、信息脱敏、少量高质量数据增强）；进行数据版本控制与详细文档化。**产出：**完全标准化、多重校验、高可信度并附有完整文档的数据集。

2.3 与字节跳动团队的协作模式

通过定期线上会议、即时通讯和共享文档平台，与字节跳动团队保持紧密协作。主要包括需求对接与澄清、进度同步与风险预警、中间成果反馈与迭代（如根据字节对样本的反馈调整对齐策略），以及最终验收与问题追踪，确保数据处理紧密围绕模型训练需求并快速响应变化。

三、 实习成果与体会

3.1 主要工作成果

- **数据交付：**深度参与三轮核心数据处理，累计交付超过 1w 条高质量对齐数据，涵盖 SFT、RLHF 等类型。
- **效率提升：**参与优化辅助系统等，使得第一轮处理效率平均提升约 45
- **质量体系贡献：**协助完善对齐标准细则，参与设计并实施了多项数据质量自动检测规则。

3.2 遇到的典型问题与解决方案

- **问题：复杂指令对齐标准理解与执行一致性。 解决方案：**组织专项讨论，细化 SOP；加强培训；引入“三人校验”；开发辅助脚本检测不一致性。

- **问题：大规模数据版本管理与效率瓶颈。 解决方案：**引入规范的数据命名与版本控制策略（如 Git LFS/DVC）；优化脚本性能（如并行处理）；利用分布式计算资源。

3.3 实习收获与感悟

此次实习使我将在校理论知识应用于 AI 前沿实践，收获颇丰：

- **专业技能提升：**深化了对 LLM 数据工程及对齐技术的理解与实践（SFT/RLHF 数据构建）；增强了 Python 数据处理与分析能力；熟练应用了相关开发工具与公司内部数据平台。
- **综合素养锻炼：**在快节奏项目中提升了团队协作、跨团队沟通、问题解决能力；培养了严谨细致的工作态度、责任心和时间管理能力。对 AI 行业发展、技术趋势及伦理挑战有了更直观的认识。

四、 总结与展望

4.1 实习总结

在文远知行为期六个月的大模型实习，是一段极具价值和挑战的成长经历。我深度参与了字节跳动“豆包”大模型的数据支持项目，通过完整地执行三轮核心数据处理与对齐流程，不仅将在校所学的理论知识与前沿的工业实践紧密结合，更在实际操作中锻炼和提升了专业技能、团队协作能力和行业认知。我深刻体会到高质量数据对于驱动 AI 发展的基石作用，也认识到大模型数据工程的复杂性和精细度。虽然过程中遇到了不少挑战，但在导师的悉心指导和团队成员的共同努力下，都得以妥善解决，并从中汲取了宝贵经验。我为能参与顶级 AI 项目贡献力量而自豪，也认识到自身在算法理解、系统优化等方面的不足，将是未来学习重点。

4.2 未来学习与职业规划展望

本次实习经历极大地激发了我对 LLM 及其数据生态的浓厚兴趣。未来计划：

- **理论深耕：**系统学习深度学习、NLP、强化学习前沿理论，特别是 LLM 训练、微调、评估相关算法。
- **技术拓展：**关注高级 LLM 对齐技术、数据治理与隐私计算、模型可解释

性及多模态大模型数据处理。

- **工程实践**：持续提升编程与工程素养，学习高效数据处理框架（Spark/Ray）和 MLOps 工具链。

职业上，期望继续在 AI 领域从事大模型算法研发、数据工程或 AI 产品相关岗位，应用实习经验，致力于构建更智能、可靠、负责任的 AI 系统，为 AI 技术发展贡献力量。