题目 2： AI 应用开发平台实践经历分享

随 AIGC 技术的日益普及，Dify、扣子等 AI 应用开发平台为开发者们提供了便捷、低代码的方式来创建、测试和发布智能体。你是否曾经在这些平台上尝试构建过自己的智能体？请分享你的实践经历，包括创建目的、过程、遇到的挑战以及最终实现的成果。

基于 Dify 平台的客诉标签 AI 实践经历分享

1. 实践背景

2024 年 7-11 月，我在好未来-用户体验组担任商业数据分析师一职。在周会中，上级对同部门负责用户研究的同事提出了对客诉文本进行分析，产出月报、季报、学期报的需求。然而，对每个月 10000-15000 条的**客诉文本预处理**是十分消耗人力和时间的，因此同事与我讨论运用 AI 进行文本数据清洗和数据分析的可行性。

以下所有内容与步骤，除特殊标注外，均由本人一人完成。

考虑到信息安全问题，无法将部分信息（例如完整的 Prompt，知识源，运行数据，运行链接等）带出好未来公司，因此在本文档中不予展示，敬请谅解！

2. 平台选择

该产品运用的是 Dify 平台（由好未来产品部提供使用），通过编排工作流、知识库 RAG 检索等方式，实现所需功能。

3. 智能体构建的实践过程

3.1.需求分析

需要处理的数据为：来自用户致电等 5 个渠道，对于科学素养等 5 种课程，包括好评和差评（建议），共计 5*5*2 种类型的文本。不同类型的文本有不同的特点，在文本长度、主题数量等方面存在差异，因此需要进行个性化的处理。数据量在每月 10000-15000 条，每条平均 100-120 个汉字，共计 100-180w 汉字。

需求如下：

- 情感分析：需要判断出文本是好评还是差评（建议），需要考虑好评低星和差评高星的情况，还需要考虑好评中存在建议和建议中存在好评的复杂情况。
- 类型分析：需要基于已经搭建好的指标体系，判断每一文本属于什么类型。例如差评文本可能是批评老师上课缺乏关注激励的，可能是老师过度催缴的，可能是课程难度偏高或偏

低等等。

后文将这两种需求统称为**“打标签”**。

3.2.智能体的开发与测试

明确了需求后，智能体的开发与测试分为以下几个阶段：

3.2.1. 流程设计与框架搭建

由于不同来源、课程的文本侧重点存在差异，用户需求与画像也存在差异，因此用户研究组生成的报告也要有相应的特点体现。所以我们在正式“打标签”之前，首先运用 Python 和问题分类器（Dify 平台提供）对文本自动进行分类。对于不同业务线、情感的文本，采用不同的 Agent 进行处理。

之所以要分如此多不同的 Agent 进行处理，是因为好评、差评、不同业务线的关注点、类型都是不同的，所建立的指标体系也有差异。如果统一成一个 Agent 来处理的话，结果的准确度会收到影响。

业务线 / 情感	科学思维	人文创作	国际素养	彼芯托管	学而思高中
好评	Agent 1			Agent 2	Agent 3
差评（建议）	Agent 4			Agent 5	Agent 6
两者均有	Agent 1+Agent 4			Agent 2+Agent 5	Agent 3+Agent 6

3.2.2. System Prompt 撰写

在开发过程中，System Prompt 的撰写是确保 AI 能够准确理解任务的关键。为了帮助 AI 更好地识别文本内容并进行准确的分类，我们采用了“立角色 + 述问题 + 定目标 + 补要求”的结构来编写 Prompt。这个结构能够清晰地传达给 AI 任务的背景、目标以及特定的要求。此外，考虑到运行成本和效率，将 Prompt 长度控制在 1500-2000 个汉字。

例如：

角色设定：你是一个 AI 文本分类助手，专门帮助分析来自不同渠道的客户反馈，并准确为每条反馈文本打上类型标签。 *（此处以类型标签举例，不对情感进行分析）*

问题描述：客户反馈的文本数据量大且内容复杂，包含多种情感和主题。我们需要你帮助我们根据预设的标签体系，对每条反馈文本进行分析，判断文本的情感倾向，并根据已定义的标签体系给出准确的类型分类。

目标设定：

你的目标是为每条文本选择一个或最多两个最相关的类型标签，标签包括：

教师质量（例如上课方式、教学内容、互动情况）

退费问题（例如退费流程、效率问题）

课程难度（例如课程过难或过易）

服务质量（例如客服、前台等）

……

具体要求：

1. 只输出标签名称，不要输出其他内容。
2. 如果情感和类型标签有多个可能，并且这些标签都同样重要，请以“标签 1,标签 2”的格式输出，标签之间用英文逗号分隔。只有当两个标签同等重要时才输出多个标签。
3. 如果存在前因后果明确的情况（例如客户投诉退费问题但最终目的是投诉教师），请分析并选择最相关的标签输出，不需要输出两个标签。

请根据以上要求进行分析并输出最相关的标签。

3.1.1. 知识库搭建与 RAG 检索

考虑到 AI 不能全面理解每个标签所对应的具体情况，我们决定通过搭建一个知识库，结合 RAG 检索（Dify 平台提供）来辅助 Agent 在没有完全理解某些概念时，从知识库中获取相关信息。这种方式能够弥补 AI 对具体标签的理解不足，提供更多上下文信息，提高其准确性。

我们选择了 RAG 检索而不是将所有信息直接写入 Prompt 中，主要是因为这样能够**降低成本**并**提高运行效率**。每次调用时，系统只需要从知识库中检索相关信息，而无需每次都阅读大量的重复内容。

3.1.2. 测试与微调参数（温度等）

在调试阶段，我们通过调整温度、TOP P、存在惩罚、频率惩罚等参数，使得生成的结果更加贴近预期。

- 温度调节：控制生成文本的随机性，温度较高时，模型生成的文本更加多样，温度较低时，生成的文本则较为确定。通过反复调试，确保情感分析能够稳定识别文本的情感倾向。我希望模型严格从已有标签中选择，而不是创造新标签，因此选择了较低的温度值（0-0.2）。
- TOP P：用于控制生成时考虑的概率分布的范围，p 值较小会让模型只选择最可能的选项。为避免模型产生不符合标签体系的内容，我们选择设置为 0.7 - 0.9。这样可以确保生成内容在较为集中的范围内，避免随机性过强，且能确保模型输出符合要求的标签。
- 惩罚机制：通过调整“存在惩罚”和“频率惩罚”来减少重复内容的生成，保证 AI 输出内

容的多样性与有效性。我们的智能体不需要对词汇的存在和频率进行惩罚，因此都选择了较小的数值（0.1-0.3）。

在确定范围后，经过具体案例的测试，最终确定温度等参数。

3.1.3. 模型选择、成本控制与效率提升

在 AI 智能体的开发过程中，模型的选择对系统的效果至关重要。我们对比了多种主流语言模型，包括 GPT-4、GPT-4o、GPT-3.5、Claude 3.5、智谱清言等。我们在综合考虑准确性、成本、效率后，最终选择了 GPT-4o 模型。

在正式上线前，我做了最后一轮的优化、精简、微调 Prompt 和模型，进一步控制运行成本，并保证了高效的处理速度。

3.1.4. 交付

在封装上线后，为了便于团队独立使用，我组织了培训，提供了详细文档，并建立了后期维护流程。确保即使我离职了之后，部门成员也能够对系统进行使用、维护、优化。

4. 遇到的挑战与解决方案

在实践过程中，我们遇到了一些技术难点和平台限制。

首先，**文本分类的准确性**是一个挑战，尤其是当文本中包含复杂、多重的情感或主题时。例如：

- 案例 1：家长写了一长段话表扬老师，但是最后说了 2 句话批评前台很冷漠。（**复杂情感**，需要分别进行识别和分析）
- 案例 2：家长因为老师上课不好，选择退费，但是由于退费效率很慢，打电话投诉。（**复杂主题**，需要准确分析家长想投诉的是老师还是退费）

为了解决这个问题，我们加强了知识库的建设，利用 RAG 检索技术为 AI 提供更多背景信息，从而提高了情感分析、类型分析的准确度。

其次，**模型的稳定性**也是一个长期挑战。尽管在封装上线前已经对模型进行了调优，但 GPT-4o 本身会不断迭代，导致我们调用的 API 也可能变化，直接影响结果的稳定性。为此，我与好未来产品部负责人进行了沟通，确保在模型更新时尽量保留旧模型，并将智能体配置过程教给几位同事，以便于后期模型更新后他们能自行优化和调整。

5. 实现成果与应用效果

最终，我们成功实现了一个能够自动进行情感分析和类型分析的智能体。该智能体能够对好未来来自 5 个渠道*5 条业务线的所有客诉文本（包括好评和差评，每月约 10000-15000 条，共计 100-180w 字）进行处理，准确判断文本的情感倾向，并根据预先设定的标签体系对文本进行分类。根据目标用户（同事）反馈，智能体的准确度和处理速度均达到了预期，解决了对客

诉文本预处理的难题，极大地提升了月度报告的生成效率。 **平均每月为部门节约 3-4 个工作日**（即 2 人同时工作 1.5-2 个工作日）。

6. 经验总结

在整段实践中，最让我印象深刻的是**准确度、成本、效率平衡**。这三者像一个不可能三角一样，时常需要在各自之间做出折中和调整。通过本次项目的实践，我更加深入地理解了 AI 在实际应用中如何发挥最大效能，也学会了如何在保证高效、低成本的同时，确保结果的准确性。