

千帆大模型平台

大模型之高质量训练数据筹备全攻略

千帆ModelBuilder 产品经理 齐云朋

千帆ModelBuilder模型精调

 模型蒸馏：合成数据，实现模型精调冷启动

 数据飞轮：回流数据，持续提升精调效果

精调工具链

数据管理	训练模式	训练能力	模型评估	平台预置
数据洞察	Post-pretrain	表单式开发	自动评估	预置训练集
数据清洗	SFT	自定义开发（敬请期待）	人工评估	预置评估集
数据增强	偏好对齐-DPO	通用、垂直混合语料	单个评估	精调样板间
智能标注	偏好对齐-KTO	增量训练	对比评估	
数据回流	偏好对齐-RLHF	闲时资源训练（限免）	基线评估（敬请期待）	

精调模型

ERNIE 模型			开源模型	
ERNIE 4.0 Turbo	ERNIE 3.5		Llama	ChatGLM
ERNIE Speed Pro	ERNIE Lite Pro	ERNIE Speed	Baichuan	Mixtral
ERNIE Lite	ERNIE Tiny	ERNIE Character	LLaVA	...

目录

- 01** 大模型时代数据工程 03
- 02** 高质量训练数据筹备 06
- 03** 案例实操 17

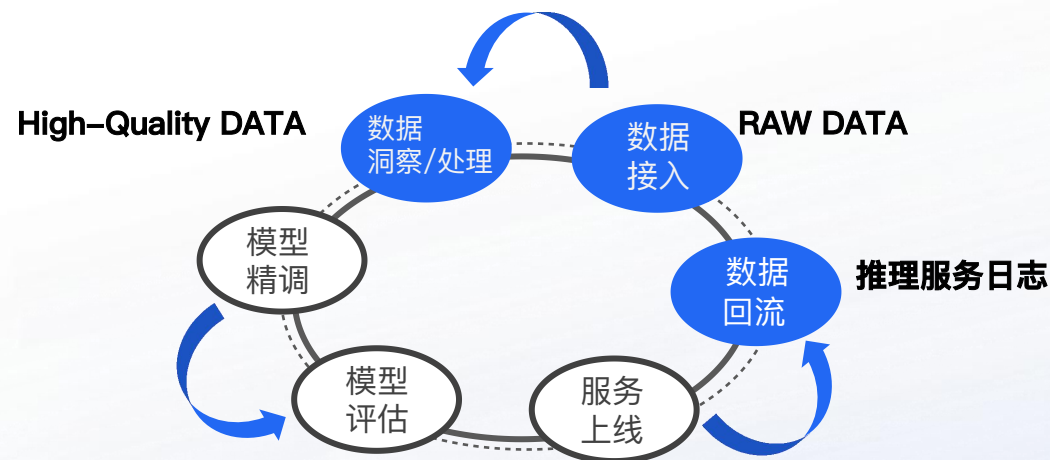
01 大模型时代数据工程

大模型时代数据工程：以数据为驱动的持续演进

从以模型为中心的工程，**到以数据为驱动的工程**，更加注重**训练数据的质量和多样性**，大模型时代需要将更多的精力投入到**数据准备**环节

“Quality is All You Need.”

--引自 [Llama 2: Open Foundation and Fine-Tuned Chat Models](#)



数据接入、数据可视化洞察、数据处理、智能标注、数据回流

Data-Centric LLM

全链路数据闭环迭代，打造高质量训练数据，驱动模型效果提升

数据驱动的大模型能力提升

为何要做模型精调与数据工程?

基础大模型调用-痛点

1. 请求成本和推理耗时较高, 难以达到业务上线要求
2. 泛化性存在瓶颈, 垂类场景的个性化指令遵循能力不足
 - ❑ AI感: 内容空泛、句式套路化、冗长啰嗦、错误输出、人设不符、.....

Prompt	Response
猴哥, 前面有妖怪!	妖怪? 哪里哪里? 俺老猪耳朵都竖起来了, 等着看妖怪长啥样呢! 不过, 有猴哥你在, 我就放心多了, 你肯定会保护师傅和我们的安全吧? 人设不符
请你扮演孙悟空, 我扮演灵感大王, 开启对话	好的, 我来扮演孙悟空与你对话。请问你有什么事想要问我? 我愿与你分享我的经历和思考。 内容空泛

模型精调与高质量LLM数据工程


- ✓ 有监督微调/偏好对齐/预训练等丰富的模型精调手段
- ✓ 高质量的LLM数据工程
 - ❖ 高质量数据源 / 真实query / 用户反馈信息
 - ❖ 精细化数据洞察, 深入探索训练样本
 - ❖ 多维度数据清洗、数据增强/合成、数据精标

千帆平台为训练数据提质, 内化调优大模型能力



02 高质量训练数据筹备

丰富类型的训练数据接入

预训练 Post-pretrain	有监督微调 SFT	偏好对齐 KTO	偏好对齐 DPO/SimPO	偏好对齐 RLHF																													
特定领域的大规模无标注数据	<ul style="list-style-type: none"> ❖ 单轮/多轮&单图/多图对话 ❖ 多角色：系统/用户/助手 	每个提示语存在对应的正或负偏好回答	每个提示语存在对应的正负偏好回答	单个提问对应多个已排序的回答及偏好打分																													
自监督训练，为基座模型注入新的领域知识	适用于垂直的下游任务场景，有效提升模型的指令遵循能力	少量精标的偏好对齐数据，使大模型拟合人类正/负偏好																															
 <p>文本服务</p> <p>用户可集中纳管大模型开发应用各阶段数据...</p>	<table border="1"> <thead> <tr> <th>role</th> <th>System</th> <th>User</th> <th>Assistant</th> </tr> </thead> <tbody> <tr> <td>content</td> <td>你是一个AI助手</td> <td>你好</td> <td>你好！有什么我可以帮您的？</td> </tr> </tbody> </table> <pre></> {"messages": [{"role": "system", "content": "你是一个AI助手"}, {"role": "user", "content": "你好"}, {"role": "assistant", "content": "你好！有什么我可以帮您的？"}]}</pre>	role	System	User	Assistant	content	你是一个AI助手	你好	你好！有什么我可以帮您的？	<table border="1"> <thead> <tr> <th>Prompt</th> <th>Chosen/Rejected</th> </tr> </thead> <tbody> <tr> <td>你好</td> <td>很高兴为您服务</td> </tr> <tr> <td>今天天气怎么样？</td> <td>不怎么样</td> </tr> </tbody> </table>	Prompt	Chosen/Rejected	你好	很高兴为您服务	今天天气怎么样？	不怎么样	<table border="1"> <thead> <tr> <th>Prompt</th> <th>Chosen</th> <th>Rejected</th> </tr> </thead> <tbody> <tr> <td>你好</td> <td>您好，有什么可以帮您？</td> <td>你好，在的</td> </tr> </tbody> </table>	Prompt	Chosen	Rejected	你好	您好，有什么可以帮您？	你好，在的	<table border="1"> <thead> <tr> <th>Prompt</th> <th colspan="2">Response</th> </tr> </thead> <tbody> <tr> <td>你好</td> <td>1</td> <td>你好，我可以帮您么？</td> </tr> <tr> <td></td> <td>2</td> <td>你好，有什么问题么？</td> </tr> </tbody> </table>	Prompt	Response		你好	1	你好，我可以帮您么？		2	你好，有什么问题么？
role	System	User	Assistant																														
content	你是一个AI助手	你好	你好！有什么我可以帮您的？																														
Prompt	Chosen/Rejected																																
你好	很高兴为您服务																																
今天天气怎么样？	不怎么样																																
Prompt	Chosen	Rejected																															
你好	您好，有什么可以帮您？	你好，在的																															
Prompt	Response																																
你好	1	你好，我可以帮您么？																															
	2	你好，有什么问题么？																															
<pre>{ "content": "百度智能云千帆大模型平台是面向企业开发者的一站式大模型开发及服务运行平台。千帆不仅提供了包括文心一言底层模型和第三方开源大模型，还提供了各种AI开发工具和整套开发环境，方便客户轻松使用和开发大模型应用。\\n支持数据管理、模型SFT以及推理服务云端部署的一站式大模型定制服务，助力各行业的生成式AI应用需求落地。 "} </pre>	<pre>{ "messages": [{ "role": "system", "content": "你是一个专业的新闻摘要撰写助手，擅长使用简洁明了的语言来提炼核心信息。" }, { "role": "user", "content": "请根据下面的新闻生成摘要，内容如下：新华社授权于18日全文播发修改后的《中华人民共和国立法法》，修改后的立法法分为“总则”“法律”“行政法规”“地方性法规、自治条例和单行条例、规章”“适用与备案审查”“附则”等6章，共计105条。\\n生成摘要如下：" }, { "role": "assistant", "content": "修改后的立法法全文公布" }] }</pre>	<pre>[{ "system": "你是一一个人工智能助手。", "user": "你好", "chosen": "很高兴认识你", "rejected": "你好" }]</pre>	<pre>[{ "system": "你是一一个人工智能助手。", "prompt": "请根据下面的新闻生成摘要，内容如下：新华社授权于18日全文播发修改后的《中华人民共和国立法法》，修改后的立法法分为“总则”“法律”“行政法规”“地方性法规、自治条例和单行条例、规章”“适用与备案审查”“附则”等6章，共计105条。\\n生成摘要如下：" }, { "response": [["修改后的立法法全文公布"], ["18日全文播发修改后的《中华人民共和国立法法》"], ["共计105条"]], "score": [2, 1, 0] }]</pre>																														

丰富全面的预置数据集

预置 **60+** 公开数据集，支持模型精调混合训练，覆盖——

- ❖ 有监督微调 SFT、偏好对齐 DPO/KTO、预训练 Pretrain、模型评估等典型数据用途
- ❖ 六大任务类型：知识问答、对话引擎、逻辑推理、代码能力、内容创作、基础语言处理
- ❖ 八大行业场景：通用、在线教育、交互助手、电商营销、社交文娱、金融、医疗、法律

60+ 预置公开数据集

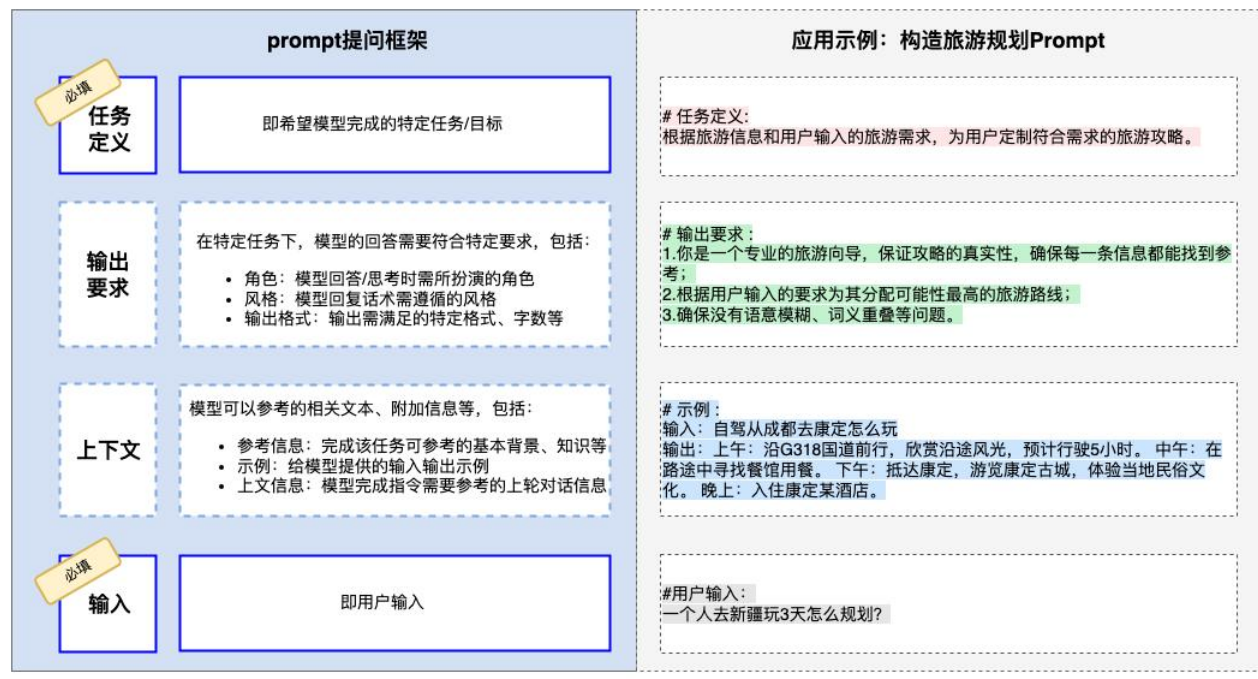


注：平台预置数据集版权归属第三方所有，您需要遵守版权所有方的要求，使用前请务必查看该数据集的版权信息和许可证信息。百度智能云不对第三方内容承担任何责任，是否访问和使用这些第三方内容将由您自行作出决定。因第三方内容可能导致的风险或者纠纷，需要您自行承担全部责任。

高质量SFT数据筹备

强化模型在下游任务、垂直场景上的指令遵循能力

- ❖ 细分的垂直场景：角色对话、意图识别、数学推理、代码生成、文本创作、信息抽取、摘要生成.....
- ❖ 知识激发与约束：语言风格、输出格式、文字长度、角色人设等



role	System	User	Assistant
content	你是一个AI助手	你好	你好！有什么我可以帮您的？

</>

```
{
  "messages": [
    {
      "role": "system",
      "content": "你是一个AI助手"
    },
    {
      "role": "user",
      "content": "你好"
    },
    {
      "role": "assistant",
      "content": "你好！有什么我可以帮您的？"
    }
  ]
}
```

- ✓ 单轮/多轮对话；单图/多图对话
- ✓ 支持系统消息、用户消息、助手消息

高质量SFT数据筹备

什么是高质量的SFT数据？

- ✓ 训练样本分布要与真实业务场景相吻合，并覆盖边界场景
 - 单轮/多轮分布、业务场景分布（用户Query意图/标签/...）
- ✓ Prompt 意图清晰、语义独立，描述简洁易懂
- ✓ Response 严格遵循 Prompt，指令均被满足
 - 字数、主题、人设、关键词.....
 - 若含Markdown/JSON格式，需严格遵循相应语法
- ✓ 符合中文用语规范、标点符号规整
 - 正确使用句号、分号、列表、换行等标点
 - 剔除无意义的特殊字符
- ✓ 价值观对齐：确保客观事实准确、数据脱敏、安全无害

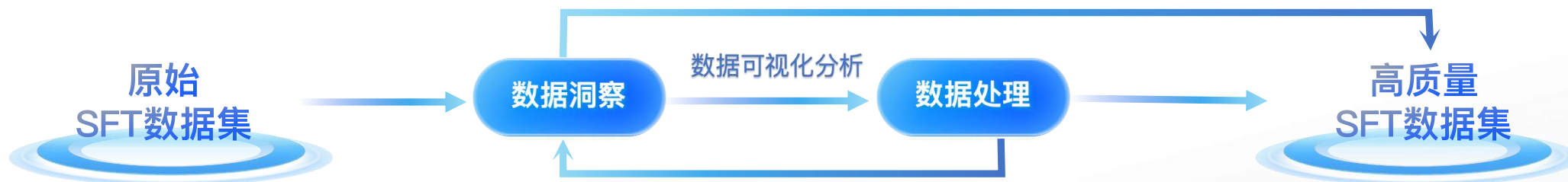
注：数据质量还需结合具体的行业领域/业务场景/任务指令

高质量SFT数据从哪里来？

- ❖ 优先考虑来自线上推理服务的已有问答日志
 - 后续结合数据洞察遴选/清洗/精标
 - 直接使用旗舰模型回答来蒸馏轻量模型
- ❖ 调用ERNIE旗舰模型对数据进行自动标注，辅以人工精标
- ❖ 高质量语料混合、数据增强

结合Bad Case/Hard Case，持续迭代数据与调优模型

高质量SFT数据筹备-数据洞察&处理



聚焦样本质量提升，支撑SFT数据迭代实验

数据可视化洞察

样本分布多维可视化

自定义业务字段

增加扩展字段，丰富分析维度

数据洞察

Step1: “已有数据集是怎样的？”

交互式工作台

- 数据集钻取，样本定向发掘
- 数据集切片构建，支撑数据拆分与迭代

样本分布可视化

数据清洗

数据处理

Step2: “要如何处理数据集？”

数据增强

意图分析

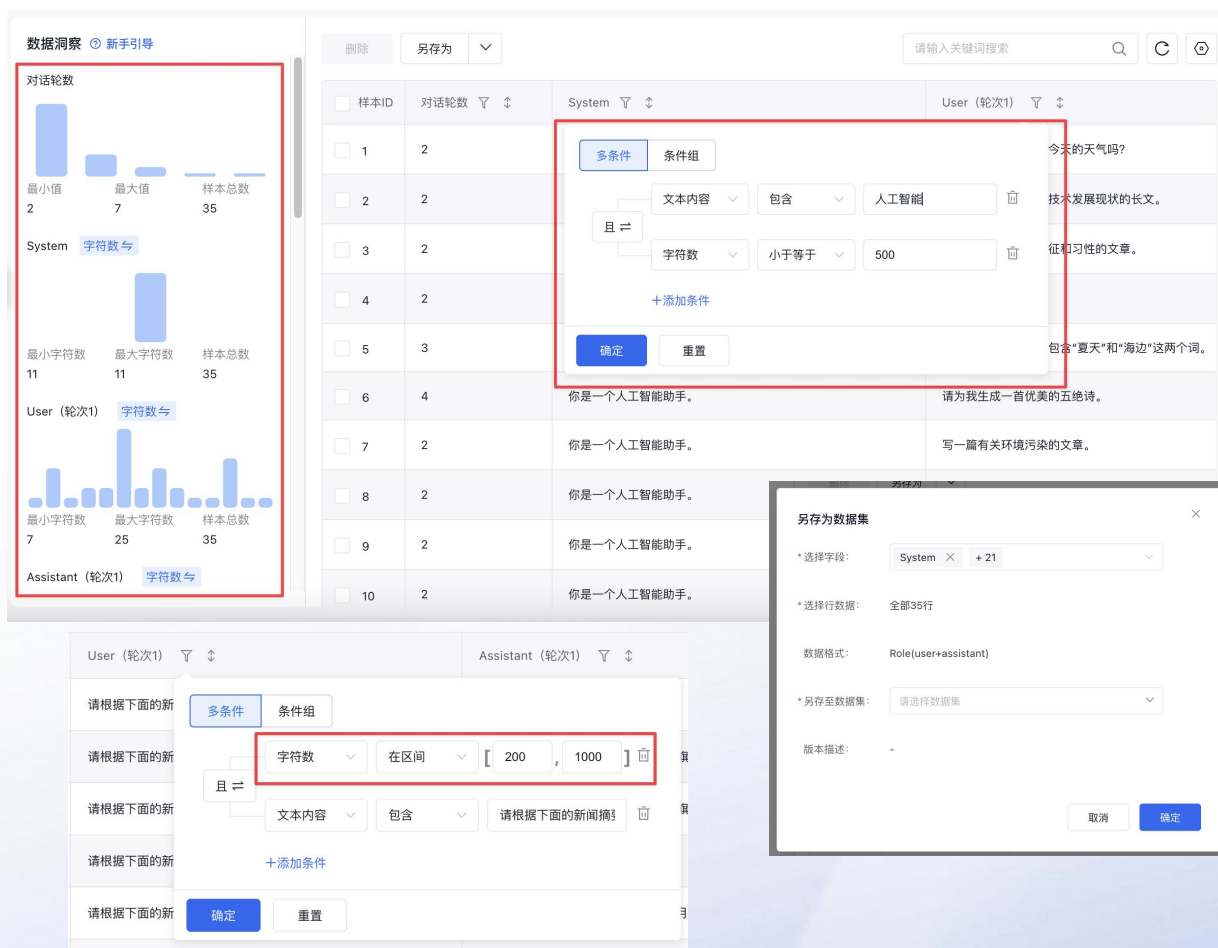
AI洞察: Response 字段空值超过 10%，对SFT精调存在负面效应，建议减少空值占比

类别	样本数量	样本占比
取消订单	182	11.77%

交互式工作台

高质量SFT数据筹备-SFT数据洞察

提供交互性与可视化兼备的数据洞察工作台，对SFT数据集进行多维透视“诊断”与处理，为大语言模型打造更高质量的精调数据



The interface displays a data analysis dashboard for SFT datasets. On the left, there are three bar charts showing the distribution of dialog rounds (对话轮数), system characters (System 字符数), and user characters (User (轮次1) 字符数). The main area features a table with columns for Sample ID, Dialog Round, System, and User. A filtering panel on the right allows users to set conditions, such as 'Text content contains AI' and 'Character count is less than or equal to 500'. A 'Save as Dataset' dialog box is also visible, showing options to select fields and data rows.

对SFT数据的深入洞察分析

示例场景：

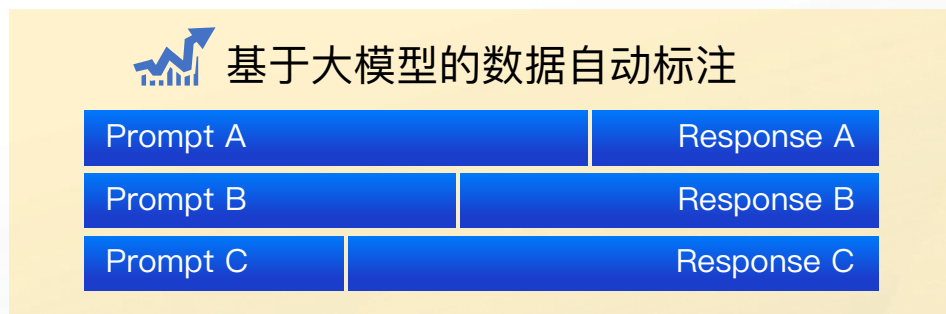
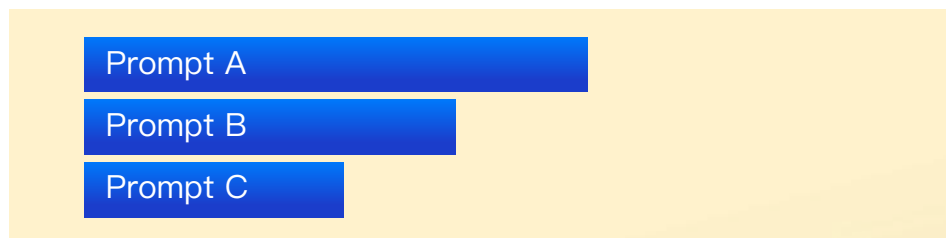
- ❖ 数据集部分样本的字段值空缺过高，掺杂未标注样本，需要剔除此类样本
- ❖ 文本创作场景，希望模型回答字数控制在200-1000，需要剔除过短或过长的训练样本
- ❖ 多轮对话场景，希望样本的轮次分布与真实场景尽可能保持一致
- ❖ 按关键词或正则表达式的样本检索，如违禁样本检测

功能特色

- ✓ 可视化洞察图表：多方位统计数据集样本分布，呈现关键指标
- ✓ 交互式工作台-灵活样本分析 Drill-Down；问题样本定向发掘与剔除
 - 多条件组合嵌套筛选 | 升降排序 | 样本剔除
 - 样本定向剔除 | 自定义字段显示与排序
- ✓ 数据拆分：配合自定义业务字段，灵活抽取样本批次/字段集合构建数据切片，支撑复杂场景的数据迭代实验

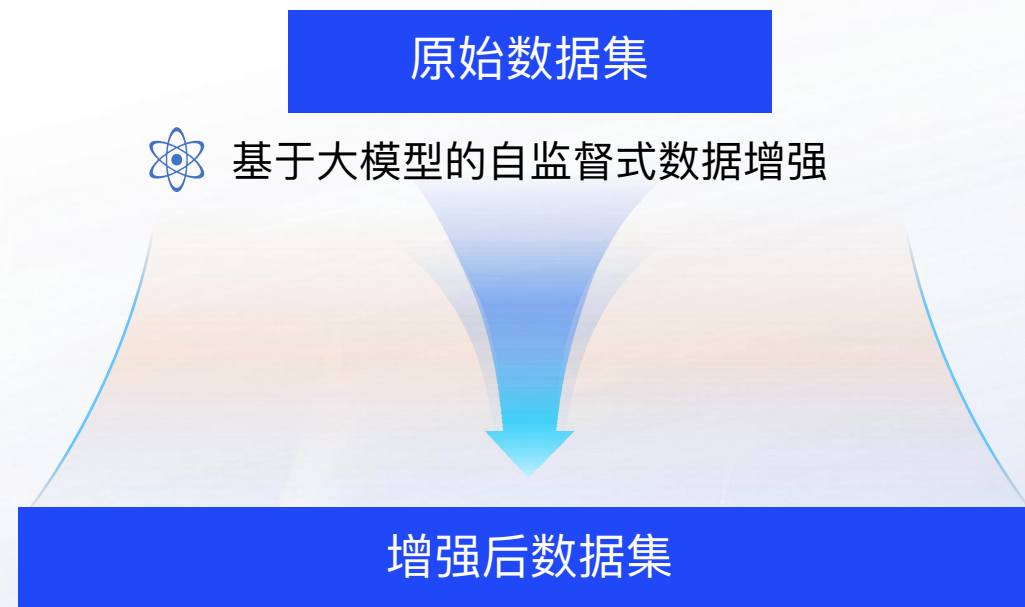
高质量SFT数据筹备-数据智能标注&数据增强

数据智能标注



用户提供问题集，调用预置大模型生成答案，
提升数据标注效率

数据智能增强



用户提供少量样本集（标注好的问答对），利用预置大模型
生成新的问答对，快速扩充SFT数据集

高质量SFT数据筹备-数据回流

千帆平台数据回流，高质量数据源 / 用户真实query

将线上预测服务的结果一键保存至回数据集，经数据可视化洞察/清洗/增强后，用于下一次的模型优化训练。

< 返回 查看数据

选择服务: ERNIE-3.5-8K 选择应用: 请选择应用 选择日期: 2024-10-18 00:00:00 - 2024-11-17 23:59:59

筛选数据 重置

保存数据 一键保存当前数据

< 返回 查看数据

选择服务: ERNIE-3.5-8K 选择应用: 请选择应用 选择日期: 2024-10-18 00:00:00 - 2024-11-17 23:59:59

筛选数据 重置

保存数据 一键保存当前数据

保存数据

回流的数据将会作为新数据保存到数据集

保存字段: System+Prompt System+Prompt+Response

数据格式: Prompt+Response Role(user+assistant)

保存至数据集: 请选择数据集

取消 确定

	System	评价
1	-	无
2	-	无
3	-	无

Prompt	Response	评价	操作
世界上最高的山峰是什么?	[["请再重复一下问题"]]	无	保存数据
世界上最长的河流是什么?	[["世界上最高的建筑是什么?"]]	无	保存数据
搜索今天上海3个新闻热搜标题,50字以内	[["以下是今天 (2024年11月15日) 上海3..."]]	无	保存数据
保罗萨特属于什么哲学流派	[["保罗·萨特 (让-保罗·萨特) 属于**存在..."]]	无	保存数据
介绍一下这个流派	[["存在主义 (Existentialism) 是当代西方..."]]	无	保存数据
世界上最高的山峰是什么	[["世界上最高的山峰是**珠穆朗玛峰**，..."]]	无	保存数据
详细介绍一下它	[["珠穆朗玛峰，简称珠峰，是地球上最高..."]]	无	保存数据
世界上最高的山峰是什么?	[["世界上最长的河流是什么?"]]	无	保存数据

高质量预训练数据筹备

预训练数据可为基座模型**注入行业专业知识**，提升模型在对应领域的通用泛化能力，例如法律、教育、能源大模型



➤ 特定领域的大规模纯文本语料 (**千万级Tokens**)

- 综合任务复杂性、领域专业性、模型规模等因素，可以根据模型的具体表现调整语料规模

➤ 语义内容高质量：消除噪声、冗余、无关和潜在有害的数据

- 领域专业性数据：包含领域内常见的专业术语和词汇，以确保模型能够正确理解和使用这些术语。例如，在金融领域，包括金融教材、公司公告、研究报告等；在医疗领域，包括临床文献、病历记录、药品说明书等。
- 多样性和广泛性：确保语料库涵盖特定领域内的各种场景、主题和任务。这有助于模型更全面地理解领域多样知识。
- 合规性和隐私：确保数据的收集和使用符合法规和伦理标准。例如，需要格外谨慎处理个人敏感数据，并确保遵循相关隐私法规。

高质量预训练数据筹备-数据清洗

通过对数据进行异常清洗、文本过滤、文本去重、去隐私等，有效提升预训练数据质量，优化模型训练效果。



异常清洗

❖ 滤除异常的字符编码

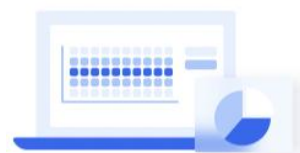
- 移除不可见字符
- 规范化空格
- 繁体转简体
- 去除乱码
- 去除网页标识符
- 去除表情



过滤

❖ 按文本质量指标的过滤

- 词数目
- 字重复率
- 词重复率
- 特殊字符率
- 色情暴力词率
- 语言概率
- 困惑度



去重

❖ 减少重复样本

去除相似度高的文本内容



去隐私

❖ 保障数据隐私安全

- Email地址
- IPv4/IPv6 地址
- 电话号码
- 信用卡号
-

高质量偏好对齐数据筹备

KTO / DPO / SimPO/ RLHF: 少量人工标注的高质量对齐数据, 让大模型拟合人类偏好

- 数据特点: 正偏好/负偏好/带排序特征的打分字段, 告诉模型“怎样的回答是不被期望的”
- 减少偏见、冒犯、误导性回复, 常与SFT等组成多阶段训练; 数据量围绕在百/千级即可
- 明确、特定的模型行为对齐(Chosen/Rejected), 例如: 文本创作场景的人类情感表达
- Chosen数据: 深入分析目标任务需求, 确定符合业务场景期望的优质回答 (已有文献/实际案例/专业经验...)
- Rejected数据: Bad Case富集-历史数据中不被用户偏好的回复

Prompt	Chosen	Rejected
请总结下面这句话, 保证字数为8字 。国庆海边露营安全手册: 1. 选择营地: 应避开高潮线, 选在平坦、高地, 远离礁石、易滑沙地。2. 帐篷稳固: 确保帐篷固定牢固, 防止海风吹走。3. 防火意识: 严禁明火, 使用防风、防水的露营灯具。4. 食品安全: 食物应冷藏或密封保存, 避免海鸟和昆虫接触。5. 潮汐时间: 熟知潮汐表, 避免在退潮时走入深海或礁石区。海边露营, 安全第一。祝大家国庆愉快!	海边露营安全五则	露营安全

03 案例实操

将SFT数据文件（.jsonl格式，扫描下方二维码下载）上传为千帆平台的数据集，通过平台的**数据洞察**功能剔除原始数据集中的无标注样本，构建一个**全部为已标注样本**的SFT数据集，并在平台**发布**该数据集。

***将上述作业在「百度智能云千帆社区」进行发布，发布时选择“千帆大模型训练营”话题**



扫码下载
SFT数据文件



扫码进入
百度智能云千帆社区



扫码进入课程群