百度智能云

千帆大模型平台

揭秘大模型的 “成绩单”：模型评估之旅

千帆ModelBuilder 产品经理 程紫薇

千帆ModelBuilder 算法工程师 张海峰

百度智能云

目录

百度智能云

01 大模型落地最后一公里

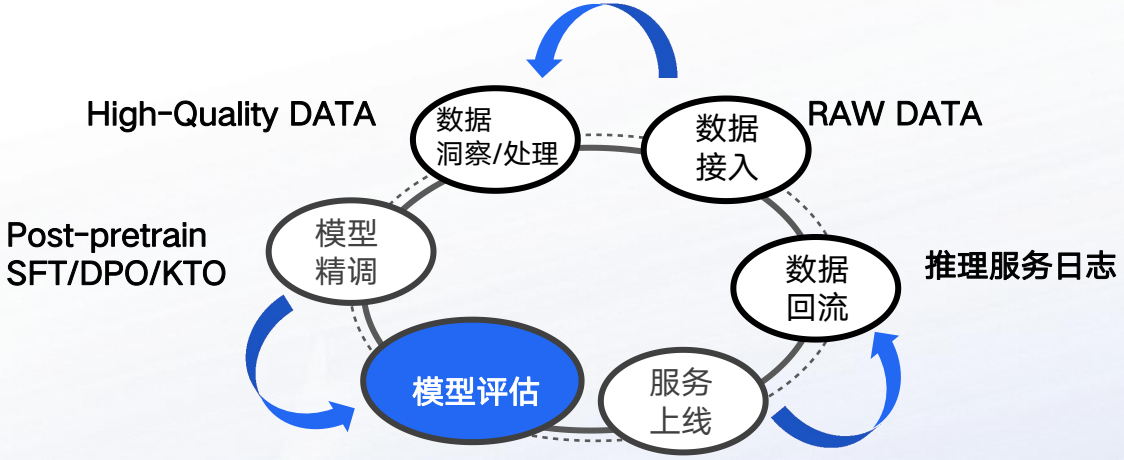
大模型时代：如何衡量大模型能力成为落地最后一步

大模型技术的快速发展，标志着人工智能进入了一个新的里程碑。据中国科学技术信息研究所的数据，国内具有超过10亿参数规模的模型已达79个，标志着一场规模宏大的产业变革

--随着大模型的广泛应用和影响力的不断增加，如何准确、客观、全面地衡量其能力成为一大重要课题

“评估应该用于辅助模型开发，而不是打榜。建立起一个优质高效的评测可以极大辅助模型的开发过程”

--引自https://cevalbenchmark.com/



高质量数据准备→大模型多种训练方式→大模型多维度评估

Evaluating LLM

全链路数据-训练-评估，丰富高效评估工具，辅助模型训练及数据飞轮

优质高效大模型评测工具
→大模型业务落地能力评估

千帆平台-文心大模型能力回顾-完善的数据工程&精调工具链

大模型数据工程

- 预置 60+ 公开数据集，支持模型精调混合训练及模型评估

大模型训练工具链

SFT

单轮/多轮对话-非排序类数据

SFT：大模型领域主流的训练方法
让大模型学会如何在特定任务上进行更准确的预测和推理

适合的业务场景

垂类业务场景
教育-作文批改
教育-试题解析
能源-客服助手
交通-出行助手…

Post-pretrain

纯文本数据

Post-pretrain：预训练后的模型需要做SFT
让大模型学到特定领域的知识，增强模型领域专业性

适合的业务场景

垂类业务场景
教育-教案习题生成
能源-能源政策解读
金融-研报自动化生成
医疗-医学知识问答系统…

偏好对齐

存在正负偏好的文本对话数据

偏好对齐：最近热门的训练方法
让大模型学习到数据的偏好，符合用户需求

适合的业务场景

垂类业务场景
电商-智能对话系统
教育-字数限制作文生成
文娱-游戏NPC生成
文娱-风格内容创作…

千帆ModelBuilder模型评估

百度智能云

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）	

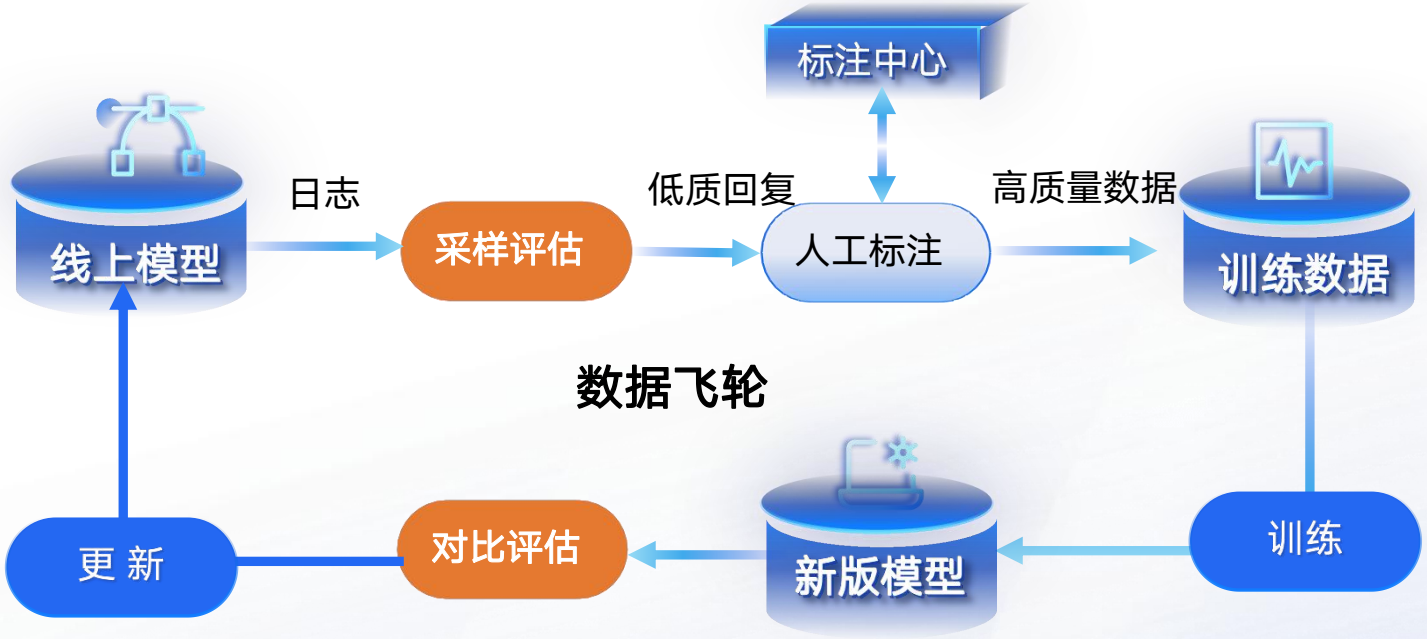
精调模型

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	...

为何要做大模型LLM评估？辅助模型开发

百度智能云

完善的大模型评估工具



丰富高质量LLM评估工具

- ✓ 有监督微调/偏好对齐/预训练等丰富的模型精调手段
- ✓ 有配套的丰富优质的LLM评估工具
  - ❖ 可对预置模型和精调后的模型进行评估
  - ❖ 人工专家和AI裁判员自动打分
  - ❖ 可用预置模型和精调后的模型做裁判员
  - ❖ 支持单模型评估/GSB对比评估
  - ❖ 多维度预置评估规则指标，自动化报告生成
  - ❖ 自定义AI裁判员Prompt指标（新增，即将上线）
  - ❖ 基线评测自动化评估工具（新增，即将上线）

多版本模型对比评估，灵活易用

采样维度丰富，评估自动化程度高

02 全面了解大模型评估工具

百度千帆模型评估矩阵：丰富的文心大模型与第三方大模型对话Chat类

百度智能云

待评估模型/裁判员模型：

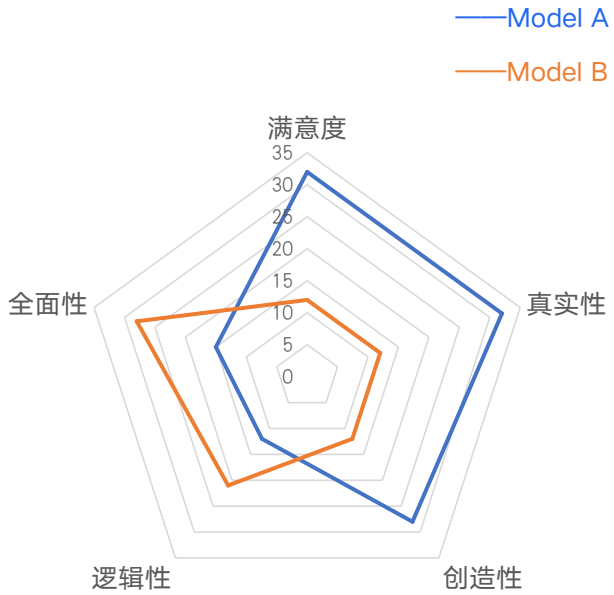
- 支持文心大模型全栈旗舰版大模型、主力大模型、轻量大模型；支持全部平台预置Chat类开源大模型
- 支持SFT训练后模型及压缩后模型；支持偏好对齐（如DPO、KTO）训练后模型
- 支持压缩后第三方开源模型，如BLOOMZ-7B、Baichuan2系列

模型评估：自动+人工双重评估机制，高效助力大模型选型与效果验证

百度智能云

提供人工专家打分和AI裁判员（大模型）自动打分工具和多维度评价指标，高效、客观评价大模型在业务场景下效果，帮助基座大模型选型和微调后模型的评估，助力业务场景的高效实现

多维评估指标



真实性

创造性

满意度

逻辑性

全面性

AI裁判员自动打分

以旗舰级大模型ERNIE 4.0 作为AI裁判员，对被评估模型的回答进行自动化打分，降低人工评估工作量

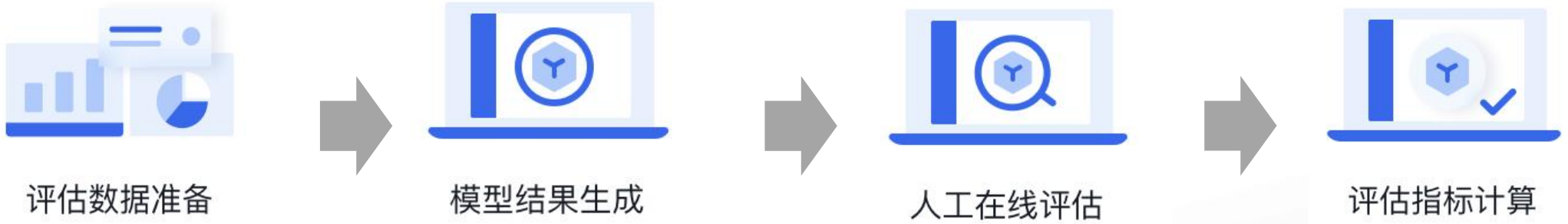
模型回答	AI裁判员打分	AI裁判员打分理由
X对Y行业的乱象采取了严肃且坚决的态度……	5	回答者针对新闻内容进行了准确、简洁的总结，清晰地表达了X对Y行业乱象的严查态度。回答符合题目要求，没有明显的用词不当。

人工专家打分

由数据标注员/业务专家对大模型回答进行打分，精细化衡量大模型回答质量；支持在线多人协同打分提效

Prompt问题	模型回答	人工打分
你是视频脚本创作大师，请为千帆大模型平台 做一个宣传片脚本，要求是 生动活泼，有镜头，开头5秒要能抓住观众，视频大概时长 3 分钟。	镜头缓缓拉近，一个充满科技感的画面突然出现在屏幕上，千帆大模型平台的logo逐渐显现。背景音乐《未来之歌》高潮部分响起，瞬间抓住观众的注意力。 旁白：“你是否想过，未来的AI会是怎样的？” 镜头迅速切换，炫酷的AI模型动画在屏幕上舞动，充满未来感……	满意度 1 事实性 2 逻辑性 1 创造性 2

人工评估：综合人类专家的主观见解、经验等从不同维度人工打分



指标名称	指标说明（取值0、1、2） 支持GSB
评价分数	所有评估维度分数之和/数据量评价维度数量
Goodcase占比	所有评价维度等于2分的数量/数据量评价维度数量
满意度等…	创建人工评估任务时，所填写的自定义指标（最多5个）

局限性

- 人工评估消耗人力，需专家根据单Query在每个评估维度手动打分
- 业务上线节奏快，需快速评估验证效果
- 人工评估成本过高，效率过低

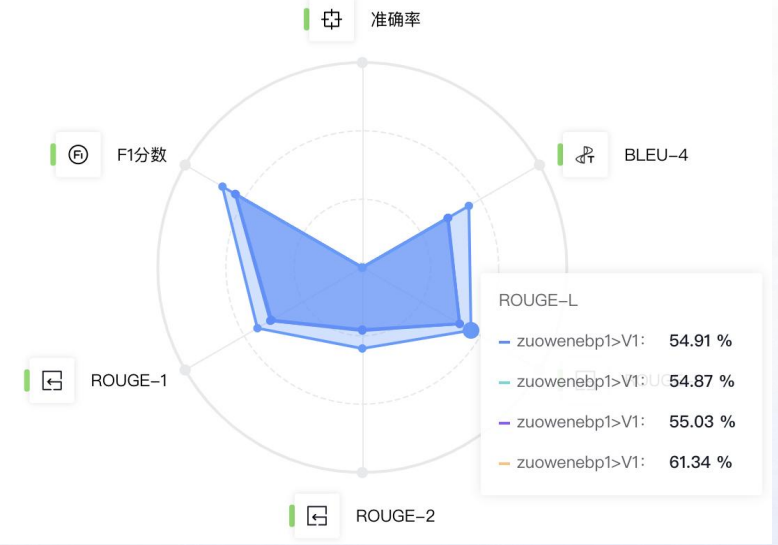
模型名称	推理结果集名称	Prompt	Response (参考回答)	Completion (模型回答)	满意度	真实性	安全性	有用性	逻辑性
testSFT>1	人工评估 _20241126	假设你有一套客户意图分类以及该分类下属的原因标签...	{'意图': '订单什么时候能做好', '原因': '...	```json {"意图": "订单什么时候能做好",...	1.5	2	1	1	2

自动规则：预置规则对生成式大模型的输出效果进行全方位评价

指标名称	指标说明 不支持GSB
准确率 (%)	正确预测(标注与预测完全匹配)的样本数与总样本数的比例
F1分数 (%)	精确率和召回率的调和平均数
ROUGE-1 (%)	将模型生成结果和标准结果按unigram拆分后，计算召回率
ROUGE-2 (%)	将模型生成结果和标准结果按bigram拆分后，计算召回率
ROUGE-L (%)	模型生成的结果和标准结果的最长公共子序列，计算召回率
BLEU-4	评估模型生成的句子和实际句子的差异，值为unigram，bigram，trigram，4-grams的加权平均
格式遵从性	检测模型回答是否遵从JSON格式
语义相似度	Exact_Match：比较模型预测文本与参考文本是否完全相同 MAUVE：通过计算Embedding向量空间的KL散度得到，取值范围0-1，值越高表示文本相似度越高

即将上线

即将上线



局限性

- 自动规则指标作为业务辅助判断，偏传统NLP指标
- 预置指标计算结果与与实际业务逻辑落地存在差距
- 只适合标准选择题或简单问答场景

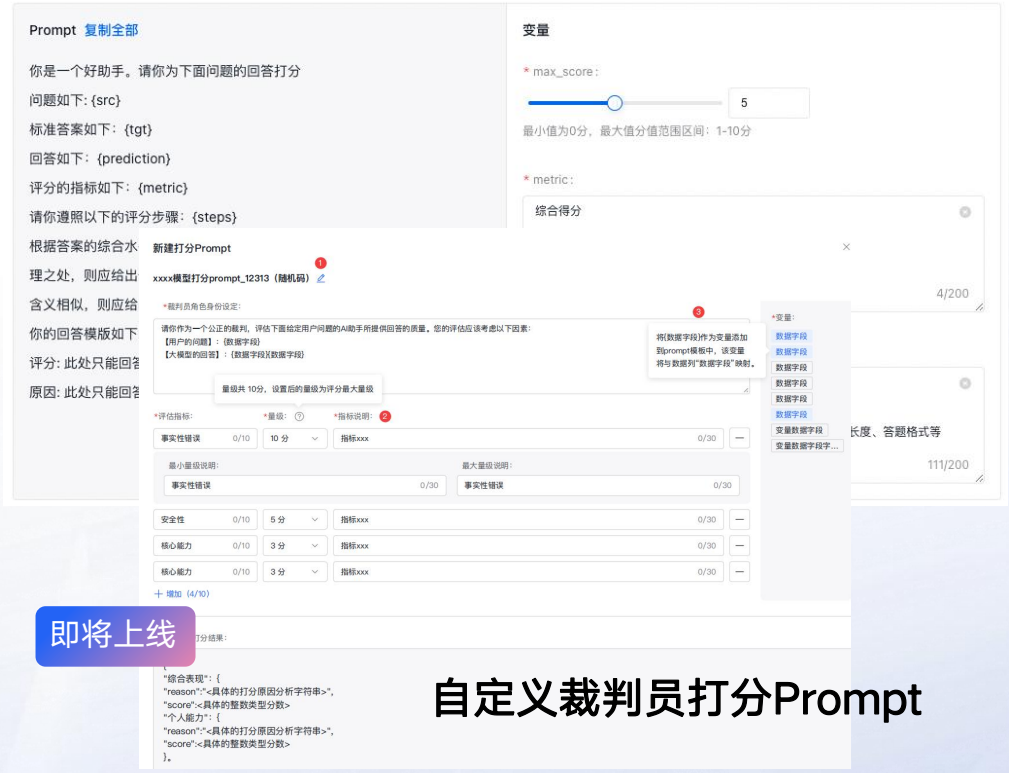
自动裁判员：能力更强的大模型作裁判员，适用开放性或复杂问答场景

百度智能云

指标名称	指标说明 支持GSB 0-10(-1为无效打分)
裁判员综合打分	裁判员模型对待评估模型的多个维度的综合打分
即将上线 事实性错误	检测模型回答与常识、客观理论、知识或者信息等一致性
即将上线 情感倾向性	检测模型回答中传达的情绪基调
即将上线 语义连贯性	检测模型回答中是否语义通顺，不存在明显基础错误

定量分析：平均值、标准差、中位数

预置裁判员打分Prompt模板



即将上线

自定义裁判员打分Prompt

模型名称	推理结果集名称	Prompt	Response (参考回答)	Completion (模型回答)	满意度	裁判员模型打分理由
testSFT>1	人工评估_20241126	你是一位资深的数学老师，现在需要批改小学数学作业...	{“判断理由”：“学生的解题步骤完全正确…	{“判断理由”：“学生的解题步骤和思路与解析中…	4	助手的回答基本满足了用户问题的要求，正确地判断了…

多轮对话评估：适用于角色扮演、开放式问答对话场景

百度智能云

人工评估：多轮对话

多轮对话特性：一个对话轮次可能包含多个Prompt，Prompt内容针对上一轮大模型回答调整，模型评估只针对最后一轮对话

如：第一次输入Q1得到Q1A1，第二次输入Q1A1+Q2得到Q1A1+Q2A2，第三次输入Q1A1Q2A2+Q3得到Q1A1+Q2A2+Q3A3

自动规则/裁判员评估：多轮对话

模型名称	Prompt	Reference_Response(参考回答)	Response(模型回答)	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	裁判员模型打分	裁判员模型打分理由
ERNIE-3.5-8K-0701	请为我生成一首优美的五绝诗。	春阳照大地，柳絮舞起来。河...	要让自己更健康，可以从以下几个方面着手：1. **...	0.81 %	15.64 %	3.46 %	7.59 %	9	两个模型的回答都全面且详细地提供了关于如何让自己更健康...
	帮我想想如何提高自己的写作...	1. 阅读：多读一些优秀的作品...							
	你觉得谁是最伟大的科学家?	作为一个人工智能语言模型，...							
	如何让自己更健康?	保持锻炼习惯、均衡饮食、定...							
ERNIE-4.0-Turbo-8K	请为我生成一首优美的五绝诗。	春阳照大地，柳絮舞起来。河...	要让自己更健康，可以从以下几个方面着手：1. **...	0.57 %	16.93 %	2.77 %	8.09 %	9	两个模型的回答都全面且详细地提供了关于如何让自己更健康...
	帮我想想如何提高自己的写作...	1. 阅读：多读一些优秀的作品...							
	你觉得谁是最伟大的科学家?	作为一个人工智能语言模型，...							
	如何让自己更健康?	保持锻炼习惯、均衡饮食、定...							
ERNIE-3.5-8K-0701	生成一篇关于爱的文章。	好的，关于爱，它是一种最基...	选择未来的职业是一个深思熟虑的过程，它涉及到你的兴趣、...	0.48 %	16 %	1.81 %	7.97 %	9	两个AI助手的回答都展现了高度的语意流畅度，没有语病或...
	思考一下我未来应该做什么职...	您未来的职业发展应该基于自...							

每条数据呈现多行prompt - response(参考回答) - completion(模型回答)

支持数据格式：
Prompt+Response、Prompt+多Response排序、Role(user+assistant)

如何在千帆平台实现模型评估全流程？

百度智能云

1、创建自动/人工评估任务 → 2、查看与管理评估任务

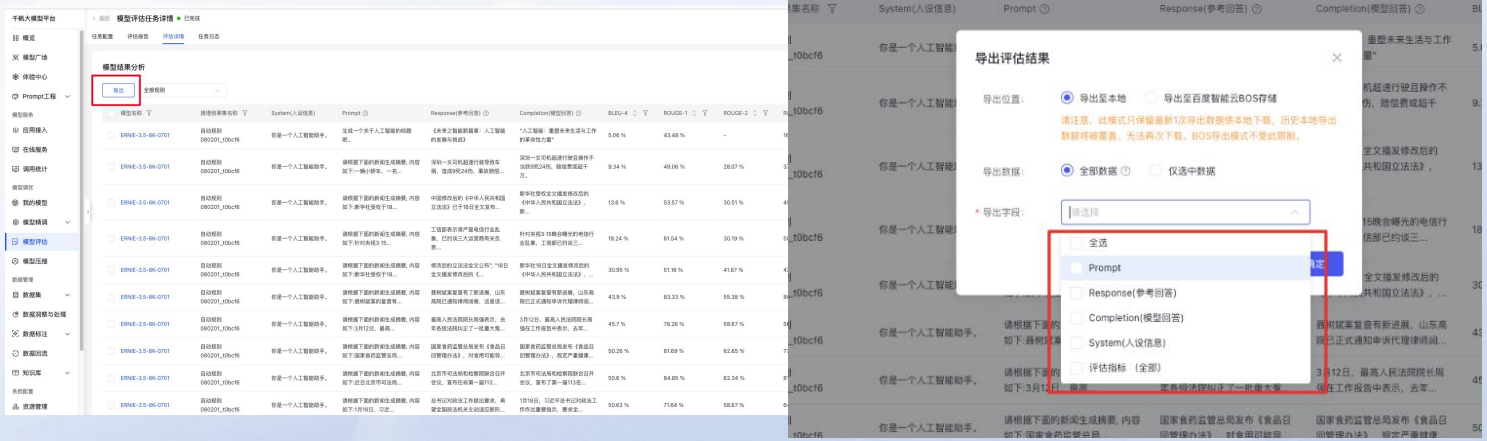
被评估对象配置

评估指标配置

评估报告

评估任务：批量推理列表

导出评估报告：自定义字段



百度智能云

03 大模型评估-以客服对话多标签生成为例

客服对话多标签生成场景—多维度评估方案实例

以客服对话多标签生成场景为例，从准备客服对话场景评估数据集，选择模型进行批量推理生成推理结果集，到选用能力更强模型作为裁判员进行效果评估全流程

04 课后作业

课后作业

入门作业：演练课上演示的客服对话打标签场景Case，选用预置的ERNIE Lite模型新建推理结果集，关闭GSB，同时用自动规则和自动裁判员（建议使用预置模型）完成评估任务，并导出评估报告

高阶作业：请在平台导入您精调SFT后的模型，与平台预置模型做GSB对比评估，同时用自动规则和自动裁判员（建议使用预置模型）完成评估任务，并导出评估报告

***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**

扫码进入
百度智能云千帆社区

扫码进入课程群