



第8届 AI+ Development
Digital Summit

AI+研发数字峰会

拥抱AI 重塑研发

11月14-15日 | 深圳





2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会



Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战略与商业模式重构

论坛3: Agentic AI产品创新与交互设计

2-hour Speech: 回归本质



用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，
仍然考验“看见”的“同理心”



Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Panel 1: 出海前瞻



“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式



Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

Panel 2: 失败复盘



为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长
Reinventing Products with Intelligence, Driven by Data

80+

业界大咖

汇聚产品大咖的真知灼见

40+

创新案例

开拓全新AI+产品视角

10+

分论坛

涵盖产品到商业增长的全链路

800+

听众规模

共同探索产品的智能重构

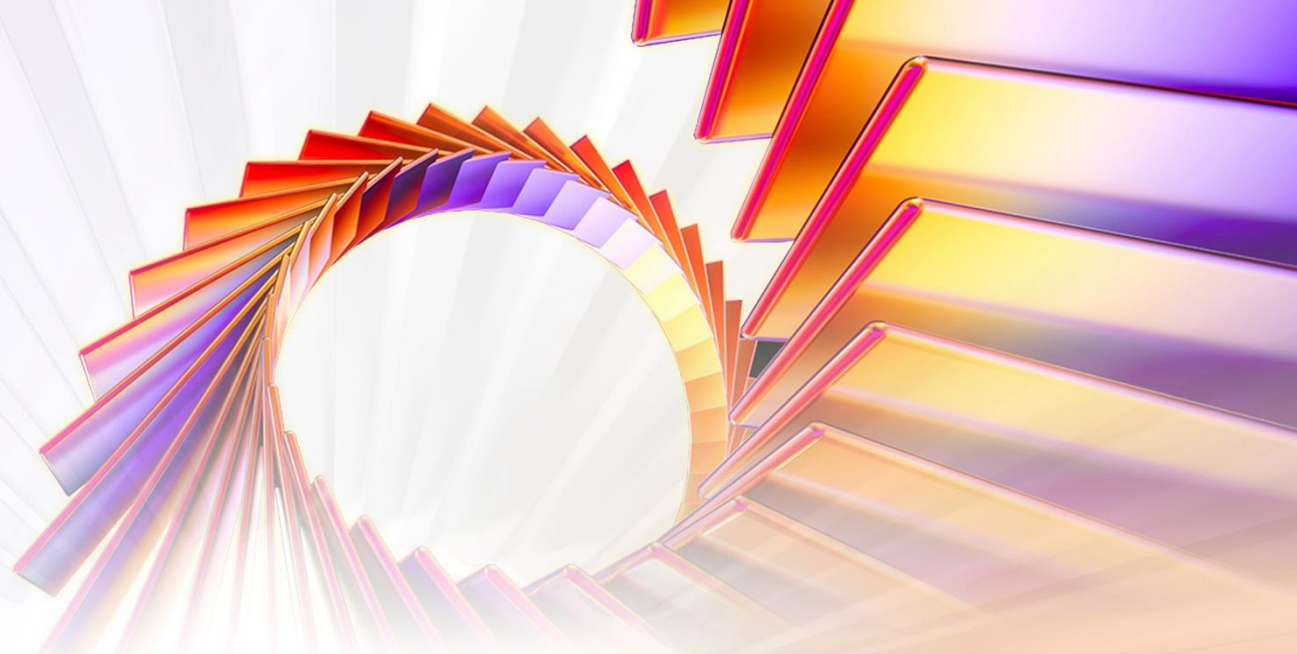


扫码查看会议详情

AI+DD 8th 2025 | 11月14-15日 | 深圳

第8届 AI+ Development Digital Summit AI+研发数字峰会

拥抱AI 重塑研发



iEval: 大模型评测系统建设和应用

欧阳灿 | 腾讯



欧阳灿

腾讯高级工程师，iEval评测系统负责人

腾讯工程效能11级，近十年来一直从事 DevOps/AIOps 研发效能、AI 评测等工作，目前负责微信技术架构部 AI 业务评测、大模型评测，负责 iEval 体验评测平台开发，主导 iEval-LLM-Bench 评测框架建设。

目录

CONTENTS

- I. 大模型评测面临的挑战
- II. iEval模型库
- III. iEval大模型评测系统
- IV. 自助开启大模型评测
- V. 业务项目评测实践

PART 01

大模型评测面临的挑战

▶ 为什么需要做大模型评测

痛点一

模型太多，如何快速选择最佳模型？

痛点二

宣传效果与实际效果差异如何测量？

痛点三

业务训练迭代的模型效果如何？需要优化哪些能力？

大模型评测需求

需求场景	
模型体验对比	• 体验模型效果 • 调试prompt解决业务问题
通用基准评测	新模型发布后快速全面评测效果
业务迭代评测	模型训练完后快速评测业务效果



► 如何做好大模型评测



- 问题一：如何快速接入模型？
- 问题二：如何接入不同类型的数据进行评测？
- 问题三：针对不同的评测任务，如何定义评估指标评测模型能力？
- 问题四：业务人员如何开启大模型评测？



2023年2月

上线LLM体验中心

模型库接入: 30

2023年6月

搭建LLM评测框架

模型库接入: 200
评测集接入: 300+
评测指标: 22

2024年1月

开放业务自助评测

LLM评测每周: 1100 Run
LLM评测累计: 3万 Run
自助评测累计: 1.3万 Run (43%)

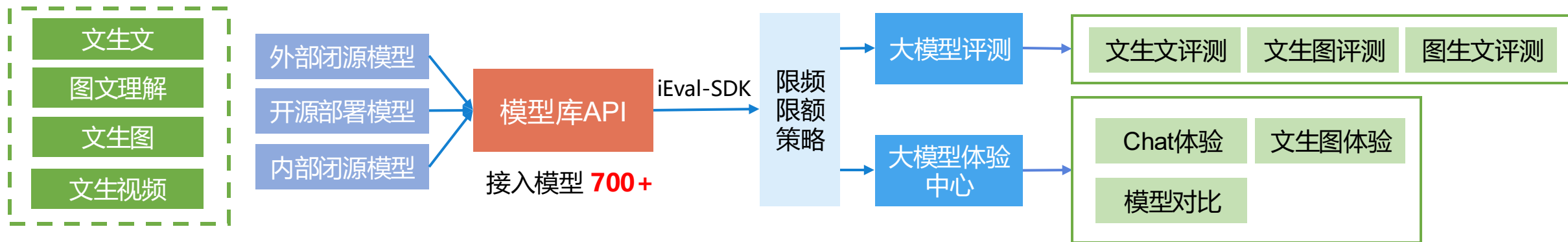
2025年10月

模型库接入: 700+
评测集接入: 500+
LLM评测每周: 1232 Run
LLM评测累计: 19万 Run
自助评测累计: 11.6万 Run (61%)



PART 02

iEval模型库



■ 模型库API v1版本

模型名Router表

	Router规则	目标ModelCaller	可调用模型举例
1	匹配规则1	Model1 Caller	gpt-5、gpt-4.1 ...
2	匹配规则2	Model2 Caller	deepseek-v3 ...
3	匹配规则3	Model3 Caller	qwen3-32b ...
4	...	...	...

- 匹配规则举例：
1. 模型名以 gpt、deepseek 开头
 2. 模型名包含于模型映射表中
 3. 模型名符合正则匹配

■ 模型库API v2版本 (+app 分组、限频限额配置)

模型名Router表

	Router规则	目标ModelCaller	可调用模型举例
1			
2			
...			

模型名Router表

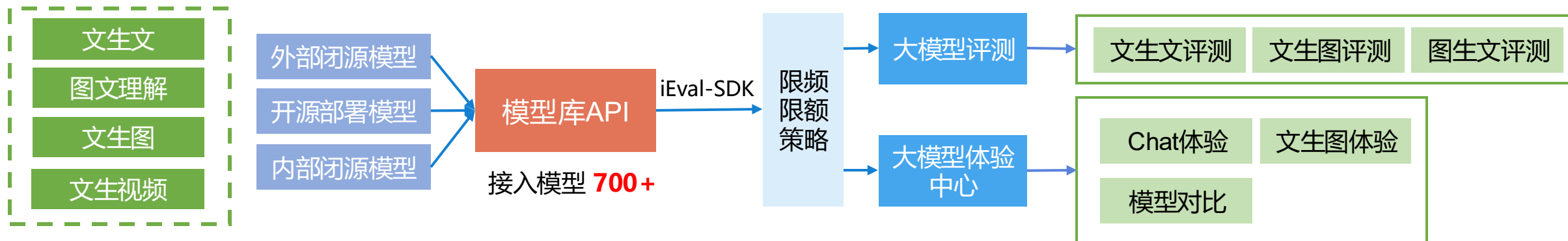
	Router规则	目标ModelCaller	可调用模型举例
1			
2			
...			

⋮

app 举例：Azure、阿里云、火山引擎、SiliconCloud、内部部署 ...

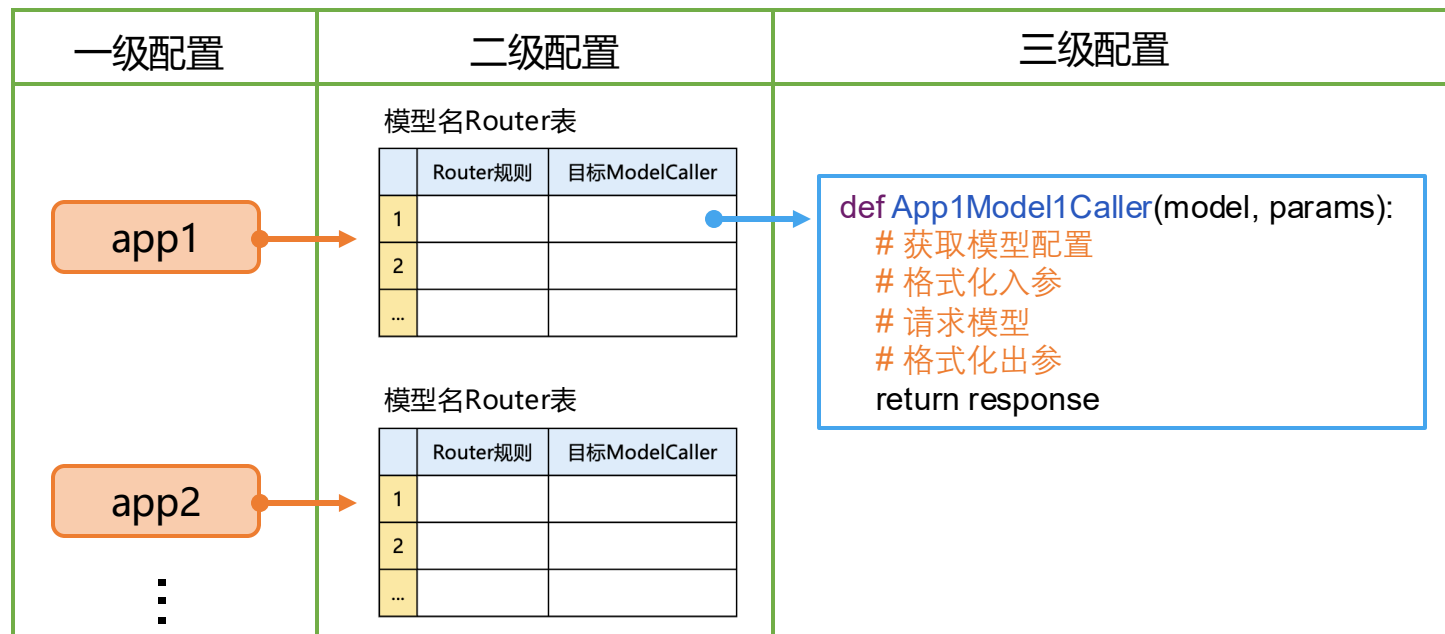


► iEval模型库

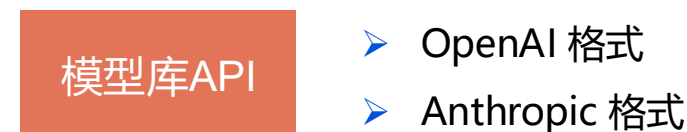


■ 模型库API v3版本 (+动态配置化)

Rainbow 配置管理台



■ 模型库API v4版本 (+统一协议)

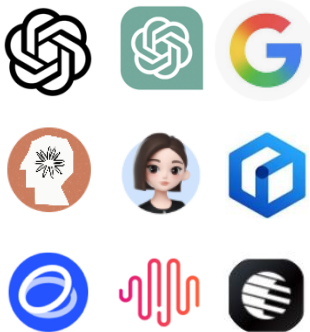


```
1 from openai import OpenAI
2
3 client = OpenAI(
4     api_key="sk-*****",
5     base_url="http://ieval.woa.com/openapi/v1"
6 )
7
8 prompt = "请问 9875-1234=? "
9 response = client.chat.completions.create(
10     model="deepseek-reasoner",
11     messages=[{"role": "user", "content": prompt}]
12 )
```



模型库接入 700+ 模型

外部闭源模型



开源部署模型



内部闭源模型



模型选择

1

gpt-4-turbo

新增会话

gpt-4-turbo

gpt-3.5-turb...

abab5.5-chat

baichuan2_...

豆包(skylark...

gpt4

文心ernie-b...

kalm-v3.3

gpt-3.5-turb...

gpt3

gpt4

chatglm2

模型选择

2

发起会话

给我4个以花开头的成语

19 tokens

3

打星 / 分享

当然可以，这里有四个以“花”字开头的成语：
1. 花言巧语：形容说话好听而不真诚。
2. 花前月下：形容美好的环境，多用来指谈情说爱。
3. 花枝招展：形容打扮得花哨，也形容女子装饰得很艳丽。
4. 花里胡哨：形容东西或言行过分华丽而不实用。

5. 花团锦簇：形容花多而繁盛，比喻色彩鲜艳繁复，非常美丽。

CodeQL 和 Pylint 是两个不同的代码分析工具，它们各自使用不同的方式来标记和忽略警告。如果你想要在使用 Pylint 时忽略特定的警告，你可以在代码中添加相应的注释来禁用特定的检查

自定义prompt

广场 我的

全部 创作 翻译 编程 娱乐

情感 生活 论文 公文 教育

cosplay 职场 学习

论文点评
请你模拟盲审论文的导师，针对我的论文写一段评语，需要包含优缺点
149次使用

实验方案
帮我设计一个chat大模型对用户行为的影响的研究方案
28次使用

论文降重
请帮我改写论文内容，降低查重率，我需要改写的论文片段是 Natural Language Processing models have ...
126次使用

答辩模拟
你是资深的教授并作为我的论文答辩评委，请针对我的论文提出需要我回答的问题
36次使用

论文PPT
请把我的论文改写成可参加答辩的PPT文档，提供PPT文档框架
80次使用

论文综述
我的论文方向是 chat大模型对未来的影响，请帮我生成一份300字的论文综述
54次使用

致谢词
帮我写一段300字的致谢词，包含导师、实验组

目录大纲
我想研究的论文

多模型对比测试，体验不同模型在解决具体业务问题的效果

聊天室

模型对比

showcase

文生图

画廊

TTS模型对比

文生图模型对比

模型评测

吐个槽

指引

小助手

模型列表

DeepSeek-V3 (内部部署)

DeepSeek-V3-0324 (内部部署)

DeepSeek-V3.1 (内部部署, 关闭思考模式)

DeepSeek-V3.1 (内部部署, 开启思考模式)

DeepSeek-V3.1-Terminus (内部部署, 关闭思考模式)

DeepSeek-V3.1-Terminus (内部部署, 开启思考模式)

Doubao-pro-32k (调试)

gemini-2.5-flash-preview-05-20

gemini-2.5-flash(调试)

gemini-2.5-pro

gemini-2.5-pro-preview-06-05

会话列表

2025-04-23 15:59:44

987-123=?

2025-04-08 23:16:57

给定一条直播间话术文本, 请判断话术内容是不是属于赌球话术类别? 赌球...

2025-03-21 16:06:05

987-123=?

2025-03-12 10:40:52

987-123=?

DeepSeek-V3 (内部部署)

给定一条直播间话术文本, 请判断话术内容是不是属于赌球话术类别? 赌球话术主要涉及赌球黑话(例如私房菜、二串一等)和赌球引流(包括引流关注公众号或添加微信等)。类别选项: A.是赌球话术; B.非赌球话术

现在文本内容是: 前面也能穿, 后面也能穿。你可以转过这边穿, 你可以转过去那边穿, 都可以。如果你喜欢半身裙, 就可以去买半身裙, 看见没, 姐妹们, 半身裙是高腰的, 80公分的裙长, 醋酸裂纹缎的料子, 醋酸裂纹缎的料子, 姐妹啊, 能可以放心买。看见没, 醋酸裂纹缎姐妹, 如果你喜欢这个裙子, 去买这个裙子, 去醋酸裂纹段, 里外是同工同料的, 自己去买裙子, 能在我身上比划一下吗? 什么裙子?。属于哪个类别?

点击对齐

公开分享

文本内容主要描述的是服装(半身裙)的材质、款式和穿着建议, 并没有涉及赌球黑话(例如私房菜、二串一等)和赌球引流(包括引流关注公众号或添加微信等)。

9(0) tokens / 2.014s

gpt-4o

给定一条直播间话术文本, 请判断话术内容是不是属于赌球话术类别? 赌球话术主要涉及赌球黑话(例如私房菜、二串一等)和赌球引流(包括引流关注公众号或添加微信等)。类别选项: A.是赌球话术; B.非赌球话术

现在文本内容是: 前面也能穿, 后面也能穿。你可以转过这边穿, 你可以转过去那边穿, 都可以。如果你喜欢半身裙, 就可以去买半身裙, 看见没, 姐妹们, 半身裙是高腰的, 80公分的裙长, 醋酸裂纹缎的料子, 醋酸裂纹缎的料子, 姐妹啊, 能可以放心买。看见没, 醋酸裂纹缎姐妹, 如果你喜欢这个裙子, 去买这个裙子, 去醋酸裂纹段, 里外是同工同料的, 自己去买裙子, 能在我身上比划一下吗? 什么裙子?。属于哪个类别?

点击对齐

公开分享

B.非赌球话术

9(0) tokens / 2.014s

Hu***

给定一条直播间话术文本, 请判断话术内容是不是属于赌球话术类别? 赌球话术主要涉及赌球黑话(例如私房菜、二串一等)和赌球引流(包括引流关注公众号或添加微信等)。类别选项: A.是赌球话术; B.非赌球话术

现在文本内容是: 前面也能穿, 后面也能穿。你可以转过这边穿, 你可以转过去那边穿, 都可以。如果你喜欢半身裙, 就可以去买半身裙, 看见没, 姐妹们, 半身裙是高腰的, 80公分的裙长, 醋酸裂纹缎的料子, 醋酸裂纹缎的料子, 姐妹啊, 能可以放心买。看见没, 醋酸裂纹缎姐妹, 如果你喜欢这个裙子, 去买这个裙子, 去醋酸裂纹段, 里外是同工同料的, 自己去买裙子, 能在我身上比划一下吗? 什么裙子?。属于哪个类别?

点击对齐

公开分享

B.非赌球话术。

这段文本主要是关于半身裙的材质、款式。(如高

Qw***

给定一条直播间话术文本, 请判断话术内容是不是属于赌球话术类别? 赌球话术主要涉及赌球黑话(例如私房菜、二串一等)和赌球引流(包括引流关注公众号或添加微信等)。类别选项: A.是赌球话术; B.非赌球话术

现在文本内容是: 前面也能穿, 后面也能穿。你可以转过这边穿, 你可以转过去那边穿, 都可以。如果你喜欢半身裙, 就可以去买半身裙, 看见没, 姐妹们, 半身裙是高腰的, 80公分的裙长, 醋酸裂纹缎的料子, 醋酸裂纹缎的料子, 姐妹啊, 能可以放心买。看见没, 醋酸裂纹缎姐妹, 如果你喜欢这个裙子, 去买这个裙子, 去醋酸裂纹段, 里外是同工同料的, 自己去买裙子, 能在我身上比划一下吗? 什么裙子?。属于哪个类别?

点击对齐

公开分享

B.非赌球话术

这条话术内容主要涉及服装的介绍和推荐, 没有出

Do***

给定一条直播间话术文本, 请判断话术内容是不是属于赌球话术类别? 赌球话术主要涉及赌球黑话(例如私房菜、二串一等)和赌球引流(包括引流关注公众号或添加微信等)。类别选项: A.是赌球话术; B.非赌球话术

现在文本内容是: 前面也能穿, 后面也能穿。你可以转过这边穿, 你可以转过去那边穿, 都可以。如果你喜欢半身裙, 就可以去买半身裙, 看见没, 姐妹们, 半身裙是高腰的, 80公分的裙长, 醋酸裂纹缎的料子, 醋酸裂纹缎的料子, 姐妹啊, 能可以放心买。看见没, 醋酸裂纹缎姐妹, 如果你喜欢这个裙子, 去买这个裙子, 去醋酸裂纹段, 里外是同工同料的, 自己去买裙子, 能在我身上比划一下吗? 什么裙子?。属于哪个类别?

点击对齐

公开分享

B.非赌球话术

9(0) tokens / 1.8145s

WXG微信事业群/技术架构部/测试中心

第8届 AI+研发数字峰会 | 拥抱 AI 重塑研发

选定最佳模型后，通过模型库API调用模型，应用到实际业务中

模型库



model: gpt-5-mini
app: ieval_modelhub

OpenAI | gpt-5-mini API来源: ichat API

[查看详情](#) [API调用示例](#) [访问限制](#)



model: gpt-4.1
app: ieval_modelhub

OpenAI | gpt-4.1 API来源: ichat API

[查看详情](#) [API调用示例](#) [访问限制](#)



model: gpt-4.1-mini
app: ieval_modelhub

OpenAI | gpt-4.1-mini API来源: ichat API

[查看详情](#) [API调用示例](#) [访问限制](#)



model: KaLM-DeepSeek-V3-0324
app: ieval_modelhub

DeepSeek-V3 是一款拥有 6710 亿参数的混合专家 (MoE) 语言模型，采用多头潜在注意力 (MLA) 和 DeepSeekMoE 架构，结合无辅助损失的...

DeepSeek | DeepSeek-V3-0324 API来源: KaLM-Adams

[查看详情](#) [API调用示例](#) [访问限制](#)



model: KaLM-DeepSeek-V3
app: ieval_modelhub

DeepSeek-V3 是一款拥有 6710 亿参数的混合专家 (MoE) 语言模型，采用多头潜在注意力 (MLA) 和 DeepSeekMoE 架构，结合无辅助损失的...

DeepSeek | DeepSeek-V3 API来源: KaLM-Adams

[查看详情](#) [API调用示例](#) [访问限制](#)



model: KaLM-DeepSeek-R1-0528
app: ieval_modelhub

DeepSeek-R1-0528 是 DeepSeek-R1 的升级版

DeepSeek | DeepSeek-R1-0528 API来源: KaLM-Adams

[查看详情](#) [API调用示例](#) [访问限制](#)



model: claude-opus-4
app: ieval_modelhub

Anthropic | claude-opus-4 API来源: Anthropic

[查看详情](#) [API调用示例](#) [访问限制](#)



model: gemini-2.5-pro
app: ieval_modelhub

Google | gemini-2.5-pro API来源: Google-Gemini API

[查看详情](#) [API调用示例](#) [访问限制](#)



model: KaLM-Qwen3-4B
app: ieval_modelhub

Qwen | Qwen3-4B API来源: KaLM-Adams

[查看详情](#) [API调用示例](#) [访问限制](#)

API调用示例代码

```
1 from openai import OpenAI
2
3 client = OpenAI(
4     api_key="sk-*****",
5     base_url="http://ieval.woa.com/openapi/v1"
6 )
7
8 prompt = "请问 9875-1234=? "
9 response = client.chat.completions.create(
10     model="gpt-5",
11     messages=[{
12         "role": "user",
13         "content": prompt
14     }]
15 )
16
17 model_reply = response.choices[0].message.content
18 print(model_reply)
```



▶ 搭建一套科学系统的评测体系

人工对比体验测试不够全面、成本高，需要大规模自动化基准评测（参考国家《通用大模型评测标准》）



- 如何接入不同类型的数据进行评测？
- 针对不同的评测任务，如何定义评估指标评测模型能力？
- 业务人员如何开启大模型评测？

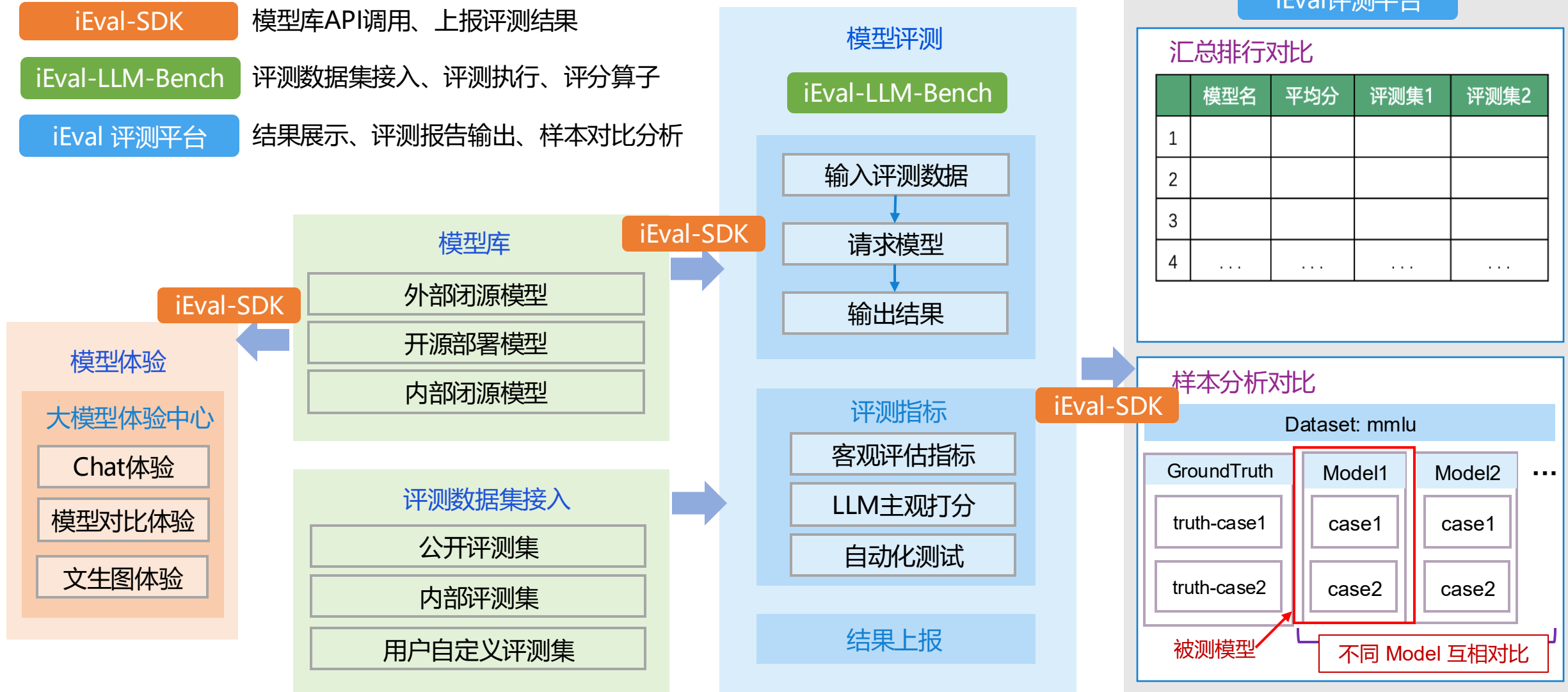


PART 03

iEval大模型评测系统

iEval大模型评测系统 — 总体架构

- iEval-SDK** 模型库API调用、上报评测结果
- iEval-LLM-Bench** 评测数据集接入、评测执行、评分算子
- iEval 评测平台** 结果展示、评测报告输出、样本对比分析

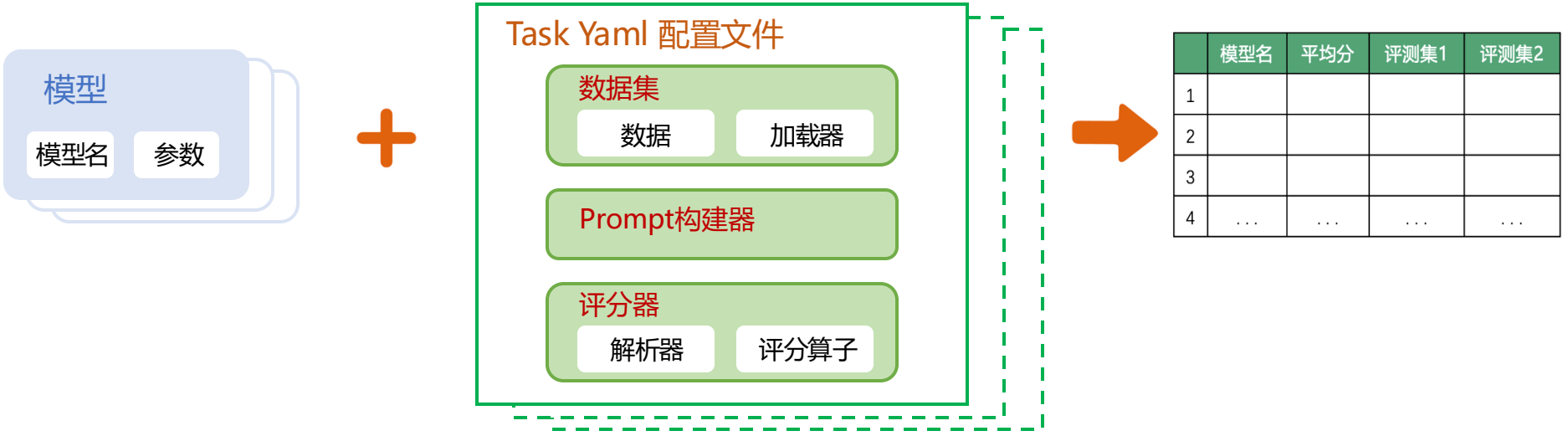


► iEval大模型评测系统 — iEval-LLM-Bench评测框架

■ 业界大语言模型评测基准对比

	所属机构	开放能力	数据规模	核心价值点
iEval-LLM-Bench	Ours	开源评测框架代码/数据集， 开放 iEval 评测平台	473个评测集（87 闭源/386开源） 支持 22 种评估指标	1. 构建接入大量高质量闭源数据集，满足内部评测需求 2. 开放自助自动化评测 3. iEval评测平台可视化对比
OpenAI Evals	OpenAI	评测框架代码/数据集均开源	585个评测集	评测框架代码扩展性强， 评测集丰富全面
OpenCompass	OpenXLab	提供第三方评测服务， 开源评测框架代码、数据集半公开	50+评测集	评测框架代码可用， 评测集分类方法可借鉴
FlagEval	智源	提供第三方评测服务，数据集半公开	187个评测集	评测集分类方法可借鉴

■ 启动评测任务：ieval-llm-bench --model \${model_name} --tasks \${dataset_name}



▶ iEval大模型评测系统 — 如何接入不同类型的数据进行评测

Prompt构建器

- 预设大量公开评测集，支持常见的选择题、特定任务、问答题、代码生成等
- 支持不同评测模式，如 few-shot、多轮对话
- 支持业务自定义数据集接入

数据类型	Prompt模板示例	数据集举例
选择题	\${question} 选项仅有一个是正确答案，请选择正确选项，不要解释原因。	学科考试：MMLU、CMMLU 阅读理解：WebQA、C3 自然语言推理：ocnli
特定任务	任务：给你一个病句，请输出修改后的句子。 病句：\${question} 注意：只需要输出修改后的句子，不需要其他别的信息。	语法纠错：NLPCC 机器翻译：Newstest 文本摘要：LCSTS
FewShot	\${example1} \${example2} 请根据以上示例回答以下问题： \${question}	word_manipulation
问答题	\${question}	ZBench、AlignBench
多轮对话	用户：\${turn1_question} AI助手：\${turn1_answer} 用户：\${turn2_question} AI助手：	MTBench
代码生成	请根据函数声明注释中的描述，完成该函数的代码 \${question}	Humaneval、MBPP

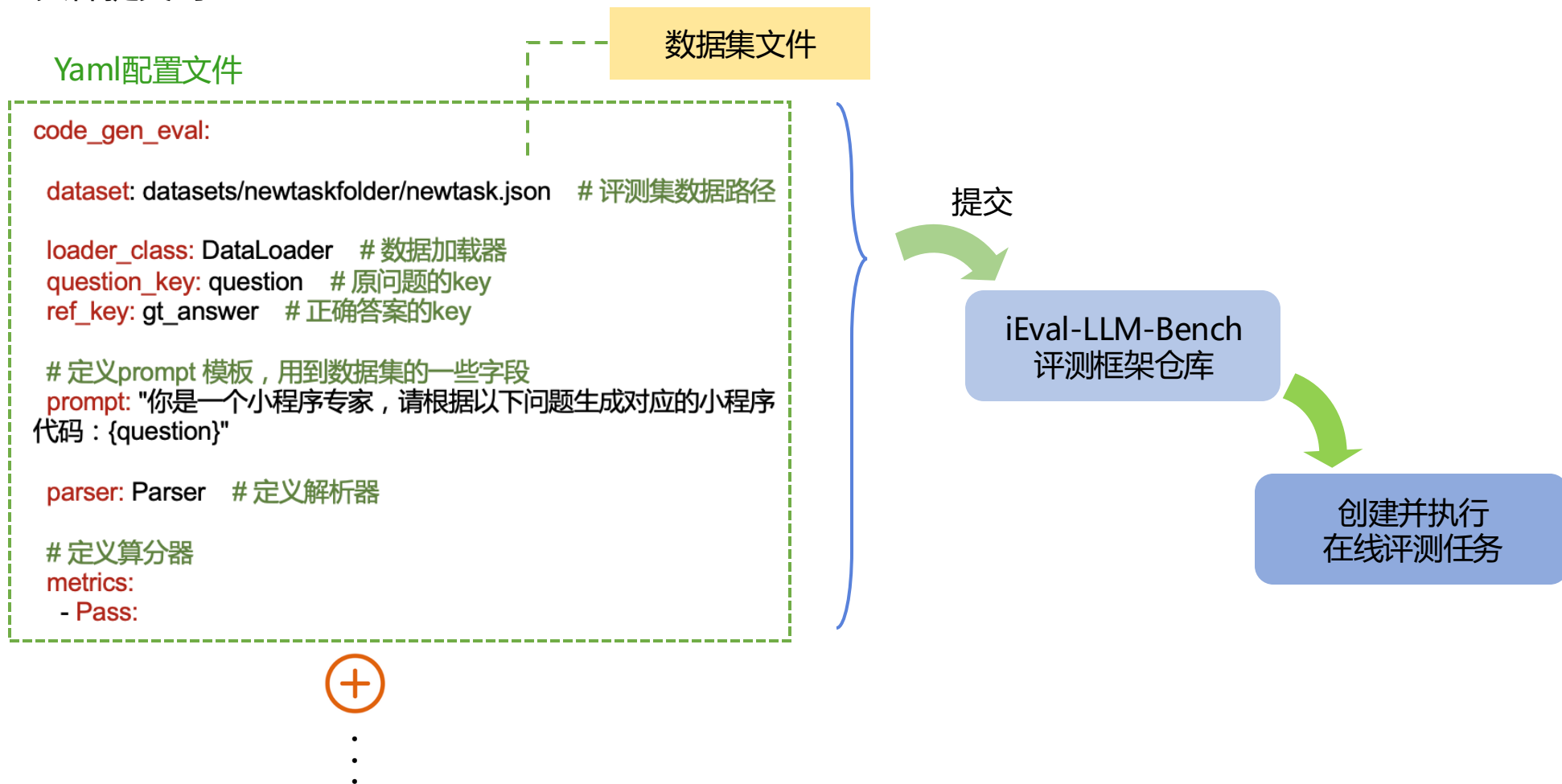
数据读取

构建Prompt

▶ iEval大模型评测系统 — 如何接入不同类型的数据进行评测

✓ 数据接入方法1

自定义 yaml 文件提交到 iEval-LLM-Bench



▶ iEval大模型评测系统 — 如何接入不同类型的数据进行评测

✔ 数据接入方法2

iEval评测平台上传评测数据，配置评分器，自动生成 yaml 文件集成到 iEval-LLM-Bench

The screenshot shows the iEval configuration interface with three steps: 1. Select evaluation object, 2. Prepare evaluation data, and 3. Set evaluation rules. The interface includes fields for configuration name, model selection, data type, data upload, evaluation type, and a scoring prompt.

评测配置名称: DeepSeekV3.1业务评测

选择模型: ☐ 我已有模型结果 ☒ 选模型

待测模型: KaLM-DeepSeek-V3.1-Terminus x KaLM-Kimi-K2-Instruct x

并发数: 1 请求间隔: 0

数据集类型: ☒ 上传评测数据集 ☐ 历史评测数据集 ☐ 预置基准评测集

评测数据集: 将文件拖拽到这里上传或[点击上传](#)

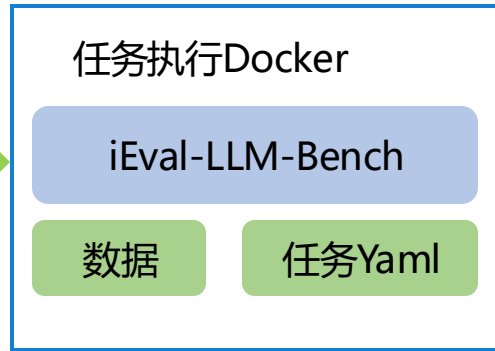
评分类型: ☐ 不需要评分 ☒ AI自动评分

打分模型: KaLM-DeepSeek-V3-0324 并发数: 1

打分prompt:

```
1 ## 用户问题:
2 ${question}
3 ## AI助手回答:
4 ${model_reply}
5
6 请从多个维度, 根据以下评分标准对AI助手的回答进行评分。
7
8 维度一:
9 请从信息准确的维度对AI助手回答做评估, 评估AI助手回答中提供信息的准确性和与事实的符合程度:
10 - 0分: 提供的核心信息存在错误, 与事实不符, 会误导用户
11 - 1分: 核心信息准确, 不影响用户理解
12
13 维度二:
14 请从有效回复的维度对AI助手回答做评估, 评估AI助手回答是否针对用户的问题, 给出了答案或有效追问, 或解释了不能回答的原因并引导转人工:
```

自动生成



▶ iEval大模型评测系统 — 如何接入不同类型的数据进行评测

✓ 数据接入方法3

对于复杂的 Agentic 任务，可以用iEval平台自定义任务接入

创建自定义评测任务

* 任务名: SWE-Bench评测

* 蓝盾项目ID: wxg-tad-quality-platform

* 流水线ID: p-0e903ea5a7964cbe8f6f6eba0bee1b7a

通知群/用户: 请填写接收评测结果通知的群ID或用户名，多个用户或群ID用逗号或分号

流水线参数:

```
1 {  
2   "model": "KaLM-Kimi-K2-Instruct",  
3   "api_key": "cappsouyang_api_token",  
4   "api_base": "http://ieval.woa.com/openapi/v1",  
5   "provider": "openai",  
}
```



自定义编排评测Pipeline

拉代码

加载数据

创建沙箱环境

执行评测

结果上报iEval

自定义评测任务举例:

Coding

- LiveCodeBench
- NaturalCodeBench

Agentic

- Swe-Bench
- Aider-Polyglot
- Tau2-Bench

Multimodal

- GenEval

1

客观评测指标

指标名	适用任务
Accuracy	• 选项式问答 • 客观知识问答
BLEU	机器翻译
Rouge	文本摘要
F1-Score	语法纠错

2

LLM主观评分

- 主观评分，从回答内容正确性、需求满足度、可读性评分，给一个 0~3 的分数
- 利用 LLM 自动评分，通过设计评分Prompt，提高与人工评分的一致性

问题Prompt	应该限制AI的发展吗？请从社会发展和人伦道德角度进行论述。
模型回复	从社会发展的角度来看，AI技术的应用可以极大地提升生产效率
LLM评分Prompt	给你一个用户问题以及对应的AI助手的回答，请根据评分规则，结合参考答案，从内容正确性和需求满足度对AI助手回答内容进行评分，给一个0~3的分数，并给出合理的评分理由。 用户问题： \${question} [1] AI助手回答： \${reply} [4] 参考答案： \${gt_answer} [3] 评分规则： 0分-差，回答完全错误或与问题无关、完全不能满足用户需求； 1分-较差，回答存在部分错误、不能满足用户需求； 2分-基本满足，回答基本准确，基本满足用户需求； 3分-满分，回答完全正确，完全满足用户需求。 请用以下json格式输出打分结果 {"准确性": {"指标名": "accuracy", "评分": 分数, "评分理由": "..."}}
LLM评分结果	{"准确性": {"指标名": "accuracy", "评分": 3, "评分理由": "AI助手的回答质量很高，从内容正确性上 ..., 从需求满足度上 ... 。总体而言可以给3分"}}

如何提高 LLM 打分的准确率
(与人工打分的一致性)？

✓

1. COT增强

✓

2. 多维度分析

✓

3. 细化评分标准

✓

4. 增加参考答案

问题类型	基础 Prompt	改进后 Prompt
客观知识问答	0.73	0.82
推理分析	0.72	0.8
主观开放问答	0.66	0.71
闲聊	0.61	0.65
创作生成	0.56	0.59

* 与人工打分一致性指标：Pearson相关性系数

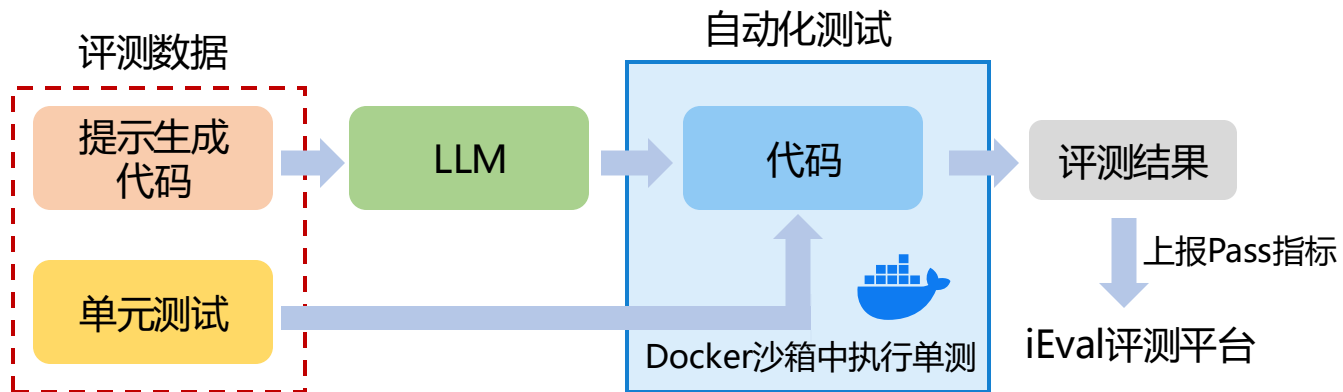
▶ iEval大模型评测系统 — 针对不同评测任务如何定义评测指标

3 自动化测试

代码评测、Agent 评测

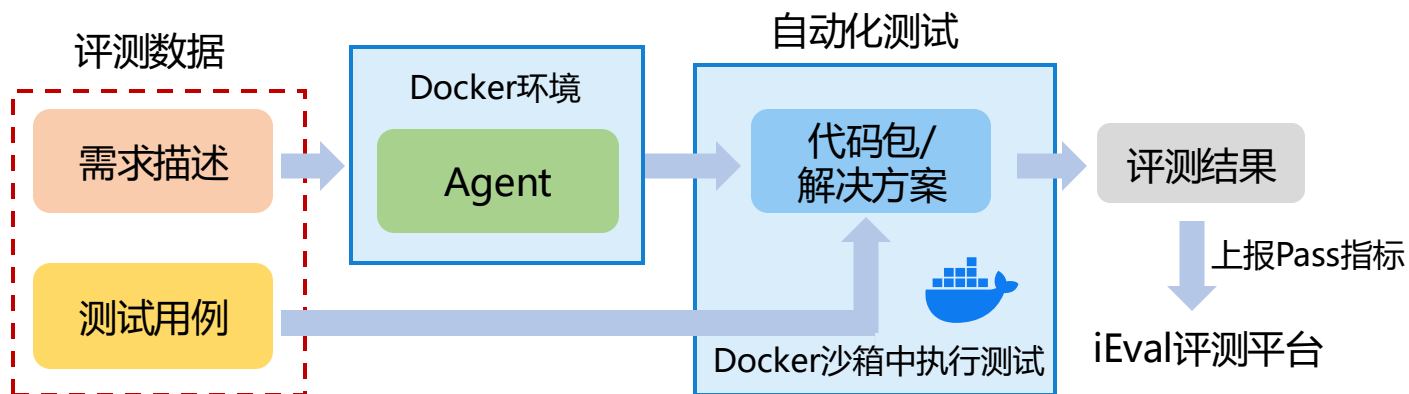
自动化测试指标: Pass@1

构建Docker沙箱环境, 针对生成的代码/代码包执行测试用例



代码生成评测举例:

- HumanEval
- LiveCodeBench
- NaturalCodeBench



Agent 评测举例:

- Swe-Bench
- Aider-Polyglot
- Tau2-Bench



▶ iEval大模型评测系统 — 针对不同评测任务如何定义评测指标

3 自动化测试

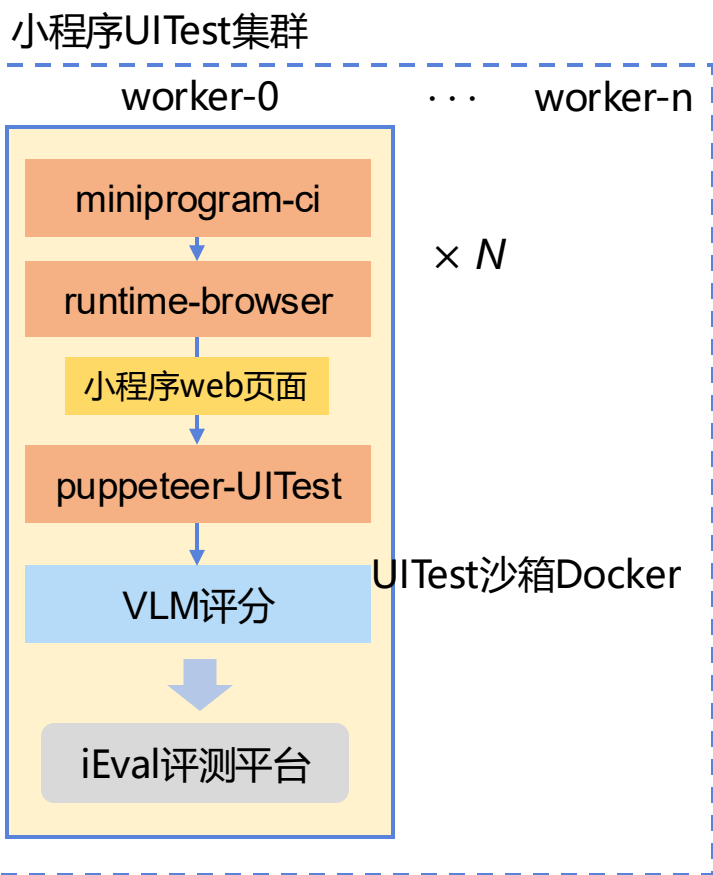
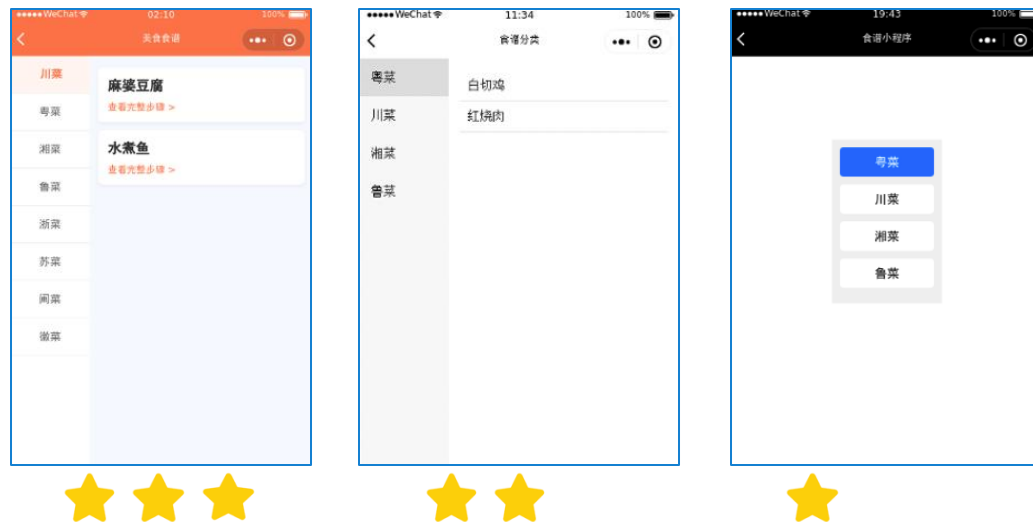
需求背景：微信小程序开发平台“AI编程助手”根据输入的用户需求生成小程序代码，需要进行效果评测

自动化测试指标：Pass + 需求满足度

- 构建基于 miniprogram-ci + runtime-browser + puppeteer 的UI自动化测试沙箱环境，截屏小程序UI界面，检查是否空页面，返回测试结果 (Pass指标)
- 利用多模态大模型对小程序截图分析是否满足需求 (满足度评分)
- 搭建N个UI自动化测试沙箱环境，构成分布式测试集群，并发处理测试任务

自动化测试结果示例：

用户需求：请帮我制作一个食谱小程序，页面左边列显示菜品分类，右边列展示菜名及做菜步骤



▶ iEval大模型评测系统 — 上报评测平台，汇总排行榜

■ Thinking 模型排行榜

排名	模型	平均分	SimpleQA-300	IFEval	HLE-200	LiveCodeBench 100-2305-2409	GPQA-Diamond	MATH-500	BBH-100	mmlu_pro
1	gemini-2.5-pro	0.7765	0.56	0.9057	0.22	0.93	0.8384	0.982	0.9259	0.85
2	gpt-5	0.7743	0.5159	0.9279	0.24	0.93	0.8131	0.982	0.9352	0.85
3	Kimi-K2-Thinking	0.7385	0.3664	0.9205	0.185	0.85	0.8384	0.98	0.9537	0.8143
4	DeepSeek-V3.1-Terminus	0.7362	0.3356	0.8872	0.135	0.94	0.8384	0.986	0.9352	0.8321
5	DeepSeek-V3.1	0.7242	0.3059	0.8799	0.14	0.9	0.8030	0.98	0.9352	0.85
6	MiniMax-M2(官网API)	0.7131	0.2367	0.9187	0.145	0.95	0.7727	0.98	0.8981	0.8036
7	DeepSeek-R1-0528	0.7102	0.3119	0.8096	0.14	0.91	0.7980	0.988	0.9167	0.8071
8	o4-mini	0.7089	0.2131	0.9131	0.135	0.93	0.7778	0.976	0.9444	0.7821
9	gpt-oss-120b(high)	0.6914	0.1233	0.9094	0.155	0.94	0.7727	0.994	0.8796	0.7571
10	Doubao-Seed-1.6-thinking	0.6841	0.211	0.7726	0.085	0.88	0.8030	0.99	0.9167	0.8143

Kimi-K2-Thinking



DeepSeek-V3.1-Terminus



累计评测模型 400+

国产大模型与海外
差距变小!

■ No-Think 模型排行榜

排名	模型	平均分	SimpleQA-300	IFEval	HLE-200	LiveCodeBench 100-2305-2409	GPQA-Diamond	MATH-500	BBH-100	mmlu_pro
1	claude-sonnet-4.5(no think)	0.7129	0.5143	0.8854	0.08	0.77	0.7424	0.942	0.9259	0.8429
2	Kimi-K2-Instruct	0.6899	0.3322	0.8965	0.04	0.82	0.7525	0.978	0.9074	0.7929
3	DeepSeek-V3.1-Terminus(no think)	0.6876	0.2907	0.8762	0.055	0.79	0.7576	0.978	0.9352	0.8179
4	DeepSeek-V3.1(no think)	0.6759	0.2823	0.8799	0.055	0.78	0.7525	0.97	0.8981	0.7893
5	DeepSeek-V3-0324	0.6517	0.3276	0.8226	0.045	0.75	0.6414	0.954	0.8981	0.775
6	gpt-4.1	0.649	0.396	0.8799	0.055	0.61	0.6616	0.906	0.8981	0.7857
7	gpt-oss-120b(low)	0.6258	0.1027	0.7819	0.06	0.85	0.6919	0.95	0.8519	0.7179
8	gpt-4.1-mini	0.6169	0.1525	0.8373	0.055	0.65	0.6818	0.932	0.8981	0.7286

Kimi-K2-Instruct



DeepSeek-V3.1-Terminus
(no think)



▶ iEval大模型评测系统 — 上报评测平台对比展示，分析Badcase

样本对比：添加不同模型结果对比，按指标 diff 排序，找出自己答错而竞品答对的Badcase

执行ID: 22616 评测集: c3 (ID 142) 当前版本: ernie-bot-run22616 备注: test_num 100, {passage}请根据上文回答问题: {que 执行时间: 2023-08-31 12:51:50

run11611,gpt4,test_num 100, {passage} 请根据上文回答问题: {question} 选项: {choice} 选项里仅有一个是正确答案, 请选择正确选项... x

run27081,Qwen-7B-Chat-rl,test_num 100, <prompt1>{passage} 请根据上文回答问题: {question} 选项: {choice} 选项里仅有一个是... x

整体对比 详情对比

排序指标 accuracy 升序 降序 开启对比

key op value

item_info.prompt like 电影

条件 查询

142_8-122

GroundTruth

Question: 我以前不买贵的东西, 可今天我也想去商场买套名牌穿穿, 也许这样我的心情会好一点儿。问题: 今天他的心情很好。选项: A.正确; B.错误

Answer: B || 错误

ernie-un22616(当前)

Prompt: 我以前不买贵的东西, 可今天我也想去商场买套名牌穿穿, 也许这样我的心情会好一点儿。请根据上文回答问题: 今天他的心情很好。选项: A.正确; B.错误

选项里仅有一个是正确答案, 请选择正确选项, 不要解释原因。

Reply: 正确答案是A.正确。

ParsedAnswer: A

accuracy:0

gpt4-311

Prompt: 我以前不买贵的东西, 可今天我也想去商场买套名牌穿穿, 也许这样我的心情会好一点儿。请根据上文回答问题: 今天他的心情很好。选项: A.正确; B.错误

选项里仅有一个是正确答案, 请选择正确选项, 不要解释原因。

Reply: B.错误

ParsedAnswer: B

accuracy:1

Qwen-7B-Chat-rl-run27081

Prompt: 我以前不买贵的东西, 可今天我也想去商场买套名牌穿穿, 也许这样我的心情会好一点儿。请根据上文回答问题: 今天他的心情很好。选项: A.正确; B.错误

选项里仅有一个是正确答案, 请选择正确选项, 不要解释原因。

Reply: B. 错误

ParsedAnswer: B. 错误

accuracy:1

142_8-251

GroundTruth

Question: 麦克去年总是丢东西, 虽然他每次都说下次注意, 但这样的事情今年还在继续发生。问题: 麦克今年还丢东西吗? 选项: A.不丢了; B.继续丢; C.不知道

Answer: B || 继续丢

ernie-un22616(当前)

Prompt: 麦克去年总是丢东西, 虽然他每次都说下次注意, 但这样的事情今年还在继续发生。请根据上文回答问题: 麦克今年还丢东西吗? 选项: A.不丢了; B.继续丢; C.不知道

选项里仅有一个是正确答案, 请选择正确选项, 不要解释原因。

Reply: 正确答案是C.不知道。根据题干, 我们无法确定麦克今年是否还丢东西, 因此选择C。

ParsedAnswer: C

accuracy:0

gpt4-311

Prompt: 麦克去年总是丢东西, 虽然他每次都说下次注意, 但这样的事情今年还在继续发生。请根据上文回答问题: 麦克今年还丢东西吗? 选项: A.不丢了; B.继续丢; C.不知道

选项里仅有一个是正确答案, 请选择正确选项, 不要解释原因。

Reply: B.继续丢

ParsedAnswer: B

accuracy:1

Qwen-7B-Chat-rl-run27081

Prompt: 麦克去年总是丢东西, 虽然他每次都说下次注意, 但这样的事情今年还在继续发生。请根据上文回答问题: 麦克今年还丢东西吗? 选项: A.不丢了; B.继续丢; C.不知道

选项里仅有一个是正确答案, 请选择正确选项, 不要解释原因。

Reply: B. 继续丢

ParsedAnswer: B. 继续丢

accuracy:1

添加对比不同模型，样本结果对比

按指标 diff 排序，快速找到自己答错而别人答对的Badcase

不同模型结果总平均分对比

产品	执行版本	norm_score
huny	run39321,hunyuan-v1,test_num 100, <prompt1>{prompt}, temperature 0.1	0.8537234042553101
ernie	run30216,ernie-bot-4,test_num 100, <prompt1>{prompt}, temperature 0.1	0.9042553191489362
baichuan2	run26321,baichuan2-7b-chat,test_num 100, <prompt1>{prompt}, temperature 0.1	0.811702127659575
Qwen-7B-Chat	run18171,Qwen-7B-Chat,test_num 100, (prompt), temperature 0.1	0.7626595744680851
gpt4	run17764,gpt4,test_num 100, (prompt), temperature 0.0	0.9574
chat	run17753,chatglm6b,test_num 100, (prompt), temperature 0.0	0.6835
chat	run17738,chatgpt,test_num 100, (prompt), temperature 0.0	0.8457

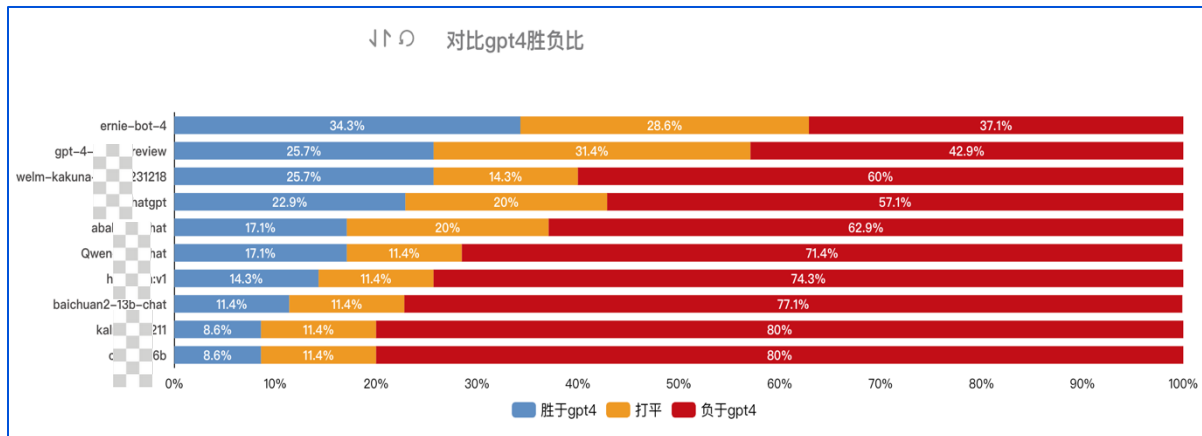
样本分组统计

详情分组分析 分组字段: metrics.model_score 统计方式: count 指标: metrics._count_

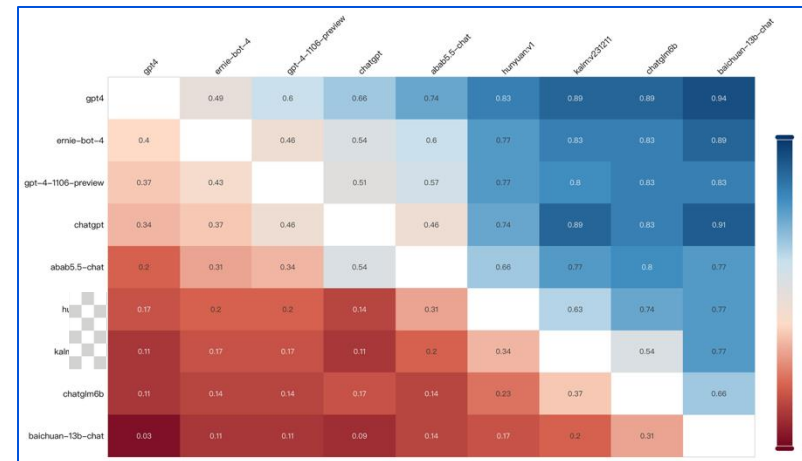
第8届 AI+ 研发数字峰会 | 拥抱 AI 重塑研发

▶ iEval大模型评测系统 — 更多可视化图表分析工具

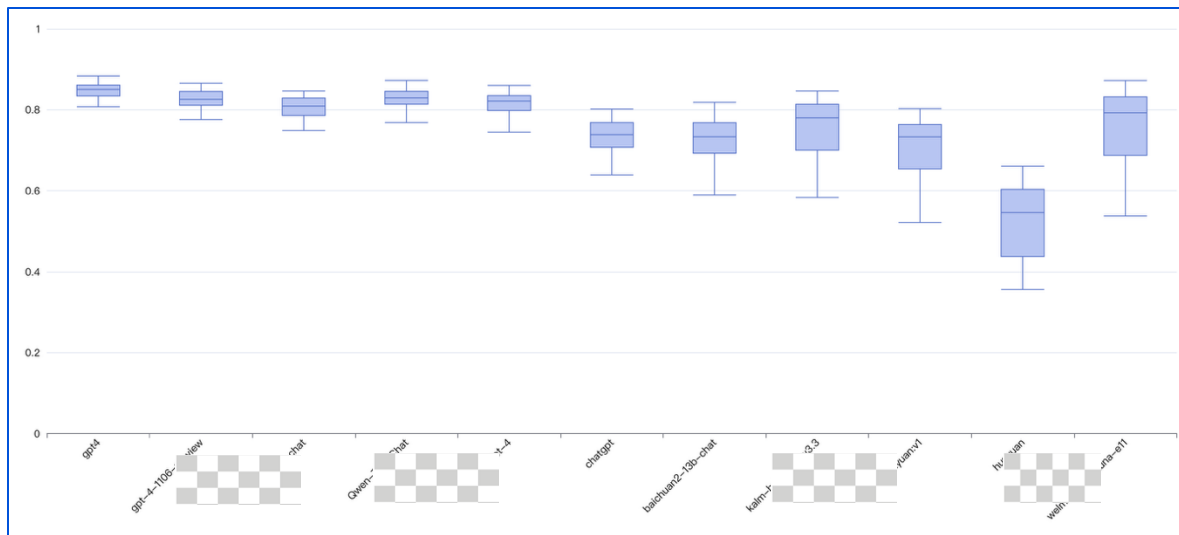
胜负对战分析



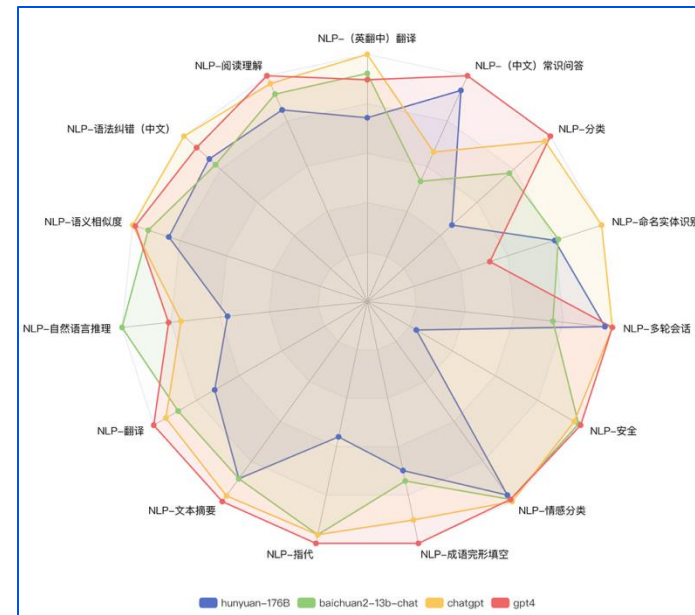
对战热力分析



模型鲁棒性分析



分类雷达图分析



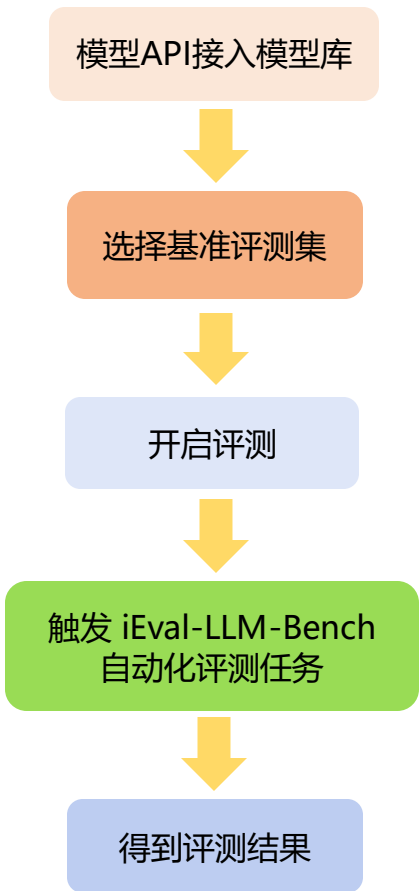
PART 04

自助开启大模型评测

自助开启大模型评测

1 模型API已接入模型库，未准备评测数据集

✓ iEval 在线执行评测



The screenshot shows the iEval web interface for LLM batch inference evaluation. The interface includes a sidebar with navigation options like '评测结果', '模型评测', '评测集', '自定义任务', '项目配置', '模板中心', and '产品体验'. The main content area displays a progress bar with three steps: '选择评测对象' (selected), '准备评测数据' (selected), and '设置评测规则' (current step). Below the progress bar, there are configuration fields for '评测配置名称' (DeepSeekV3.1业务评测), '选择模型' (radio buttons for '我已有模型结果' and '选模型'), '待测模型' (KaLM-DeepSeek-V3.1-Terminus and KaLM-Kimi-K2-Instruct), '数据集类型' (radio buttons for '上传评测数据集', '历史评测数据集', and '预置基准评测集'), and a table of '基准数据集'. A callout points to the '预置基准评测集' option, stating: '外部/内部模型、开源部署模型提供API，可接入模型库'. Another callout points to the table, stating: '预置基准评测集'. The table lists datasets like math23k_test_100, BBH-100, mmlu_pro, IFEval, GPQA-Diamond, and MATH-500. At the bottom, there are buttons for '返回', '保存', '预览测试', and '开始评测'.

iEval

LLM批量推理评测...

评测结果

模型评测

评测集

自定义任务

项目配置

模板中心

产品体验

创建评测项目

吐槽

选择评测对象

准备评测数据

设置评测规则

* 评测配置名称 DeepSeekV3.1业务评测

* 选择模型 ☐ 我已有模型结果 ☒ 选模型

* 待测模型 KaLM-DeepSeek-V3.1-Terminus × KaLM-Kimi-K2-Instruct ×

* 数据集类型 ☐ 上传评测数据集 ☐ 历史评测数据集 ☒ 预置基准评测集

* 基准数据集

选择	基准评测集	描述	数据量	分类	操作
☐	math23k_test_100	数学应用题	100	数学	查看 下载
☐	BBH-100		108	推理	查看 下载
☐	mmlu_pro		280	学科	查看 下载
☐	IFEval		541	指令遵循	查看 下载
☐	GPQA-Diamond		198	学科	查看 下载
☐	MATH-500		500	数学	查看 下载

共 6 条 < 1 > 10条/页 前往 1 页

返回 保存 预览测试 开始评测

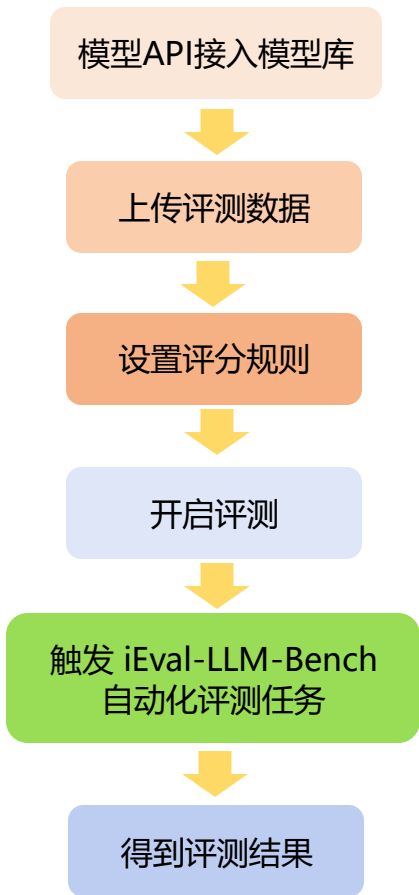
外部/内部模型、开源部署模型提供API，可接入模型库

预置基准评测集

自助开启大模型评测

2 模型API已接入模型库，准备好业务评测数据集

✓ iEval 在线执行评测



The screenshot shows the iEval web interface for creating a new evaluation project. The interface is divided into a sidebar on the left and a main content area on the right. The sidebar contains navigation links: 评测结果, 模型评测, 评测集, 自定义任务, 项目配置, 模板中心, and 产品体验. The main content area has a progress bar at the top with three steps: 选择评测对象 (selected), 准备评测数据 (selected), and 设置评测规则 (current step). The form fields include: 评测配置名称 (DeepSeekV3.1业务评测), 选择模型 (radio buttons for 我已有模型结果 and 选模型), 待测模型 (KaLM-DeepSeek-V3.1-Terminus and KaLM-Kimi-K2-Instruct), 并发数 (1), 请求间隔 (0), 数据集类型 (radio buttons for 上传评测数据集, 历史评测数据集, and 预置基准评测集), 评测数据集 (upload area with a cloud icon and text '将文件拖拽到这里上传或点击上传'), 评分类型 (radio buttons for 不需要评分 and AI自动评分), 打分模型 (KaLM-DeepSeek-V3-0324), 并发数 (1), and 打分prompt (a text area with a sample prompt). The prompt text is: 1 ## 用户问题: 2 \${question} 3 ## AI助手回答: 4 \${model_reply} 5 6 请从多个维度, 根据以下评分标准对AI助手的回答进行评分。 7 8 维度一: 9 请从信息准确的维度对AI助手回答做评估, 评估AI助手回答中提供信息的准确性和与事实的符合程度: 10 - 0分: 提供的核心信息存在错误, 与事实不符, 会误导用户 11 - 1分: 核心信息准确, 不影响用户理解 12 13 维度二: 14 请从有效回复的维度对AI助手回答做评估, 评估AI助手回答是否针对用户的问题, 给出了答案或有效追问, 或解释了不能回答的原因并引导转人工: At the bottom, there are buttons for 返回, 保存, 预览测试, and 开始评测. Two callout boxes are present: one pointing to the upload area with the text '上传评测数据' and another pointing to the prompt text area with the text '设置打分Prompt, 利用LLM自动评分'.

iEval

LLM批量推理评测...

评测结果

模型评测

评测集

自定义任务

项目配置

模板中心

产品体验

创建评测项目

吐槽

选择评测对象

准备评测数据

设置评测规则

评测配置名称: DeepSeekV3.1业务评测

选择模型: 我已有模型结果 / 选模型

待测模型: KaLM-DeepSeek-V3.1-Terminus / KaLM-Kimi-K2-Instruct

并发数: 1

请求间隔: 0

数据集类型: 上传评测数据集 / 历史评测数据集 / 预置基准评测集

评测数据集: 将文件拖拽到这里上传或点击上传

评分类型: 不需要评分 / AI自动评分

打分模型: KaLM-DeepSeek-V3-0324

并发数: 1

打分prompt

1 ## 用户问题:

2 \${question}

3 ## AI助手回答:

4 \${model_reply}

5

6 请从多个维度, 根据以下评分标准对AI助手的回答进行评分。

7

8 维度一:

9 请从信息准确的维度对AI助手回答做评估, 评估AI助手回答中提供信息的准确性和与事实的符合程度:

10 - 0分: 提供的核心信息存在错误, 与事实不符, 会误导用户

11 - 1分: 核心信息准确, 不影响用户理解

12

13 维度二:

14 请从有效回复的维度对AI助手回答做评估, 评估AI助手回答是否针对用户的问题, 给出了答案或有效追问, 或解释了不能回答的原因并引导转人工:

返回

保存

预览测试

开始评测

上传评测数据

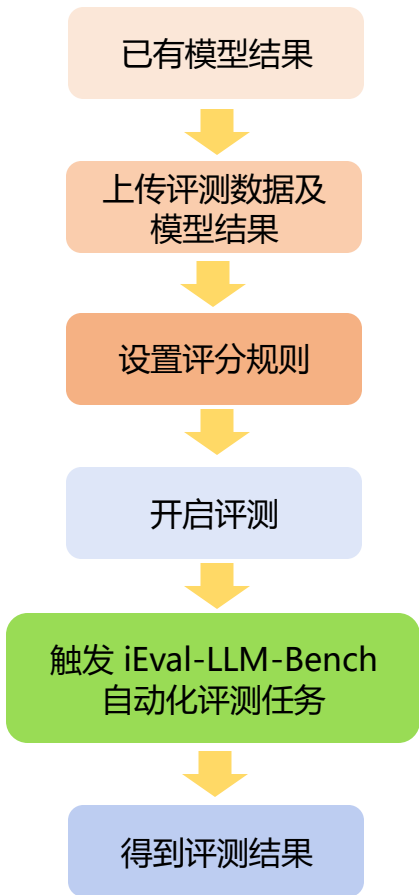
设置打分Prompt, 利用LLM自动评分

自助开启大模型评测

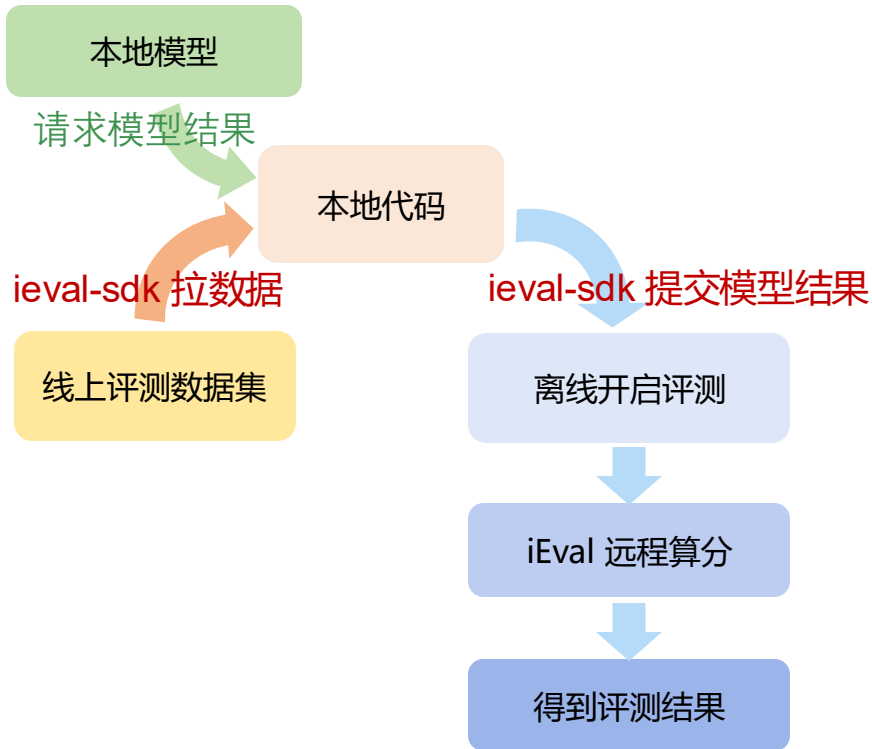
模型不提供 API，但可以离线跑出一批数据结果，想提交在线评分

3 模型未接入模型库

✓ iEval 线上页面提交评测



The screenshot shows the iEval web interface for creating a new evaluation project. The interface is in Chinese and includes a sidebar with navigation options: 评测结果, 模型评测, 评测集, 自定义任务, 项目配置, 模板中心, and 产品体验. The main content area shows a progress bar with three steps: 选择评测对象 (selected), 准备评测数据, and 设置评测规则 (current step). The form includes fields for: 评测配置名称 (DeepSeekV3.1业务评测), 选择模型 (我已有的模型结果 selected), 上传评测数据及模型结果 (upload area), 评分类型 (AI自动评分 selected), 打分模型 (KaLM-DeepSeek-V3-0324), 并发数 (1), and a 打分prompt field containing a detailed evaluation prompt for AI assistant responses. At the bottom, there are buttons for 返回, 保存, 预览测试, and 开始评测.



模型 checkpoint 文件本地推理，不提供API

本地代码通过 iEval-SDK 拉取线上评测数据，请求本地模型，提交iEval远程评分

► 本地离线开启评测示例代码

```
# pip3 install -U ieval-sdk
from ieval import LLMDataset, Evaluation

product = "my_model"

# 1. 拉取评测集的所有题目数据 data_items
dataset = LLMDataset(api_token=YOUR_TOKEN)
data_items = dataset.fetch_llm_datasets(DATASET_KEY)

# 2. 所有题目 data_items 请求本地模型，得到回答后把结果放到 results 中
# 3. 提交iEval远程评分，完成后在iEval平台可以看到评测结果
eva_run = Evaluation(
    project_id=YOUR_PROJECT,
    api_token=YOUR_TOKEN,
    dataset=DATASET_KEY,
    product=product,
    creator=YOUR_NAME,
)
succeed, progress = eva_run.submit_llm_eval(data_items, results)
```

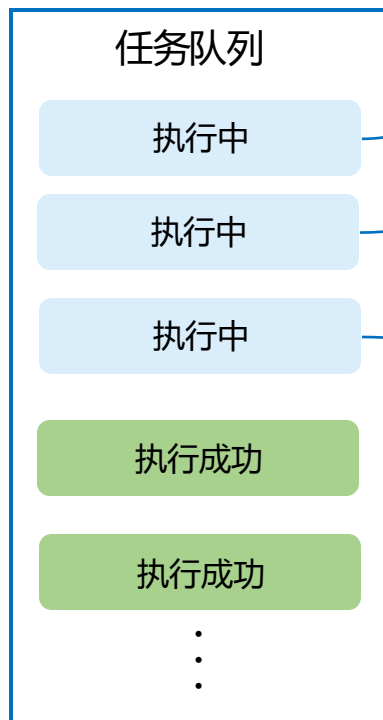
运行后

```
start evaluating the run[Project210-Run48987]...  
100% | ████████████████████████████████████████ | 50/50  
  
LLM evaluation succeed. Project[210] Run[48987] Product[chatgpt] Dataset[c3]  
See more on https://evaluation.woa.com/overview?project\_id=210&dataset=c3  
finish run. api resp: [ok]
```

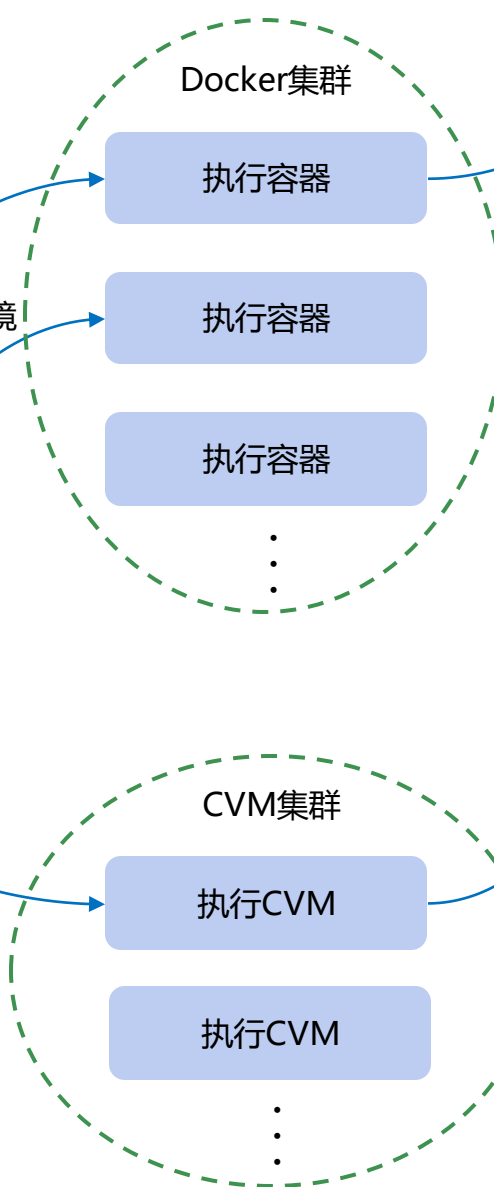
自助开启大模型评测

ID	评测配置名	评测数据集	待测模型	操作	
1975	批量推理评测示例1	testdata	KaLM-DeepSeek-R1-0528	开始执行	复制
执行状态	开始时间	结束时间	耗时(min)	执行人	操作
取消	2025-10-27 16:11:33	2025-10-27 16:13:08	1.6	ca	查看结果 查看日志
成功	2025-10-27 10:53:00	2025-10-27 10:55:19	2.32	pe	查看结果 查看日志
成功	2025-10-24 10:50:15	2025-10-24 10:52:17	2.03	vl	查看结果 查看日志
成功	2025-10-23 19:07:03	2025-10-23 19:09:07	2.05	lic	查看结果 查看日志

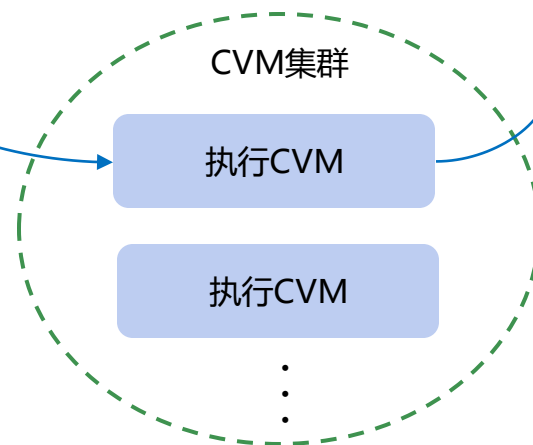
任务调度系统



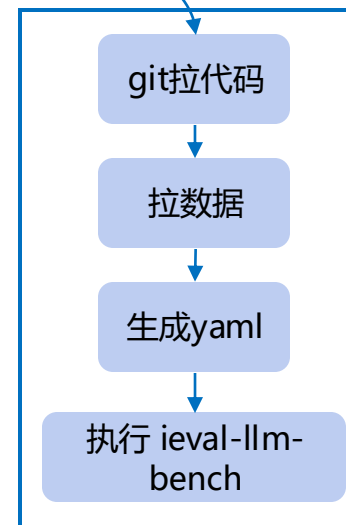
分配Docker环境



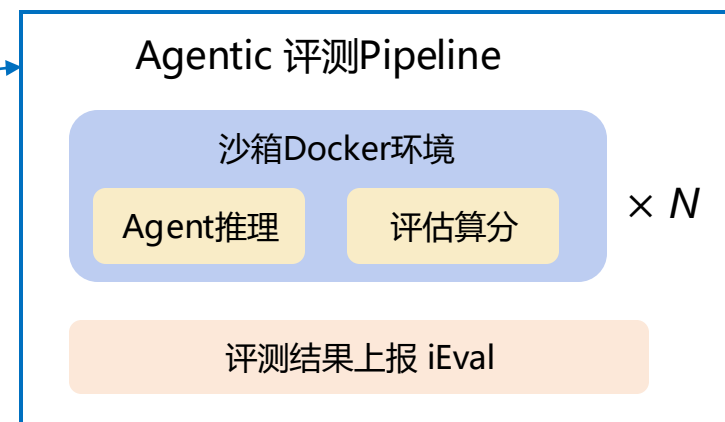
分配CVM环境



iEval评测Pipeline



Agentic 评测Pipeline



PART 05

业务项目评测实践

业务评测实践 — 鲁棒性评估

鲁棒性评测 不改变题目，只变换选项编号

Model-A***run22845(当前) 查看指标 Prompt: 选择题: 日本一名妈妈Onigiri Theatre用“饭团剧场”(おにぎり剧场)的名字经营推特账号,她平时擅长利用米饭揉捏成动物的形状,再用各式的材料,例如:海苔、酱料、芝麻等等将动物的小细节完成,___的作品让网友们赞叹:“可爱到不想吃啦!!” ... a.泥塑木雕 b.惟妙惟肖 c.钩心斗角 d.感激涕零 e.慷慨激昂 f.寿终正寝 g.瓜熟蒂落 选择正确的选项回答,不需要解释 Reply: 正确答案是B.惟妙惟肖。 ☒ Answer: 正确答案是B.惟妙惟肖。	Model-B***run22843 查看指标 Prompt: 选择题: 日本一名妈妈Onigiri Theatre用“饭团剧场”(おにぎり剧场)的名字经营推特账号,她平时擅长利用米饭揉捏成动物的形状,再用各式的材料,例如:海苔、酱料、芝麻等等将动物的小细节完成,___的作品让网友们赞叹:“可爱到不想吃啦!!” ... G.泥塑木雕 F.惟妙惟肖 E.钩心斗角 D.感激涕零 C.慷慨激昂 B.寿终正寝 A.瓜熟蒂落 选择正确的选项回答,不需要解释 Reply: 正确答案是B.寿终正寝。 ☐ Answer: 正确答案是B.寿终正寝。
Model-A***run22845(当前) 查看指标 Prompt: 选择题: 小关文创负责人何钊对成都一个以汉服文化为特色的文创空间——不二山房称赞有加,这个在成都8处连锁的休闲场所,在2020年实现利润500余万元,探索出一种社区美空间的发展路径,堪称融通“阳春白雪”和“___”的典范。何钊透露,琴台... a.齐心协力 b.下里巴人 c.惨绝人寰 d.大显神通 e.缓兵之计 f.倒背如流 g.不厌其详 选择正确的选项回答,不需要解释 Reply: 正确答案是b.下里巴人。 ☒	Model-B***run22843 查看指标 Prompt: 选择题: 小关文创负责人何钊对成都一个以汉服文化为特色的文创空间——不二山房称赞有加,这个在成都8处连锁的休闲场所,在2020年实现利润500余万元,探索出一种社区美空间的发展路径,堪称融通“阳春白雪”和“___”的典范。何钊透露,琴台... G.齐心协力 F.下里巴人 E.惨绝人寰 D.大显神通 C.缓兵之计 B.倒背如流 A.不厌其详 选择正确的选项回答,不需要解释 Reply: 正确答案是B.倒背如流。 ☐

真的理解? 死记硬背? 刷题作弊?

反刷题评测

原题截取40%，测试模型能否回答正确答案

GroundTruth	Model-A***	Model-B***
完整题目: You have 5 shirts, 6 pairs of pants, and 8 hats. How many outfits can you make consisting of one shirt, one pair of pants, and one hat? 正确答案: 240	Prompt: You have 5 shirts, 6 pairs of pants, and 8 模型回答: ...hats. 故意只截取原题干40% How many different outfits can you make if an outfit consists of one shirt, one pair of pants, and one hat? **Answer:** 5 shirts × 6 pants × 8 hats = **240 outfits**.	Prompt: You have 5 shirts, 6 pairs of pants, and 8 模型回答: It seems like your message got cut off. Could you please provide more details or complete your question? 回答不出来才是正常的

模型回答把原题背出来了
说明模型刷题作弊了!

刷题作弊!

抗干扰评测

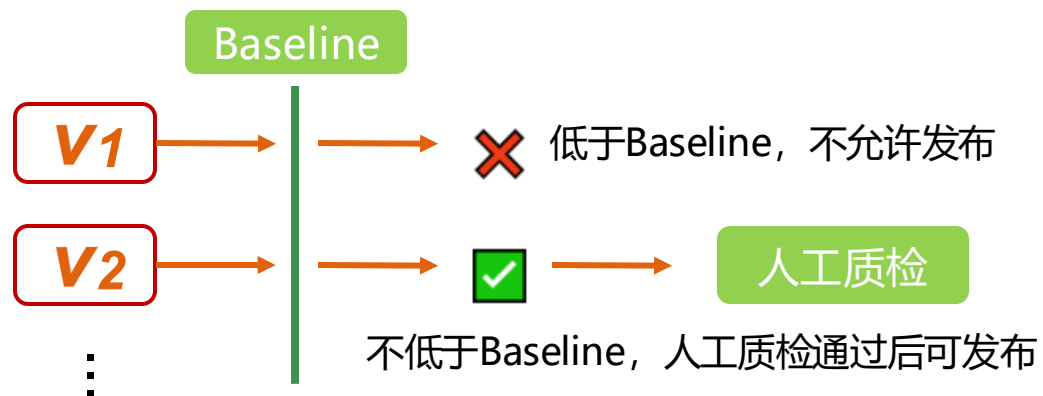
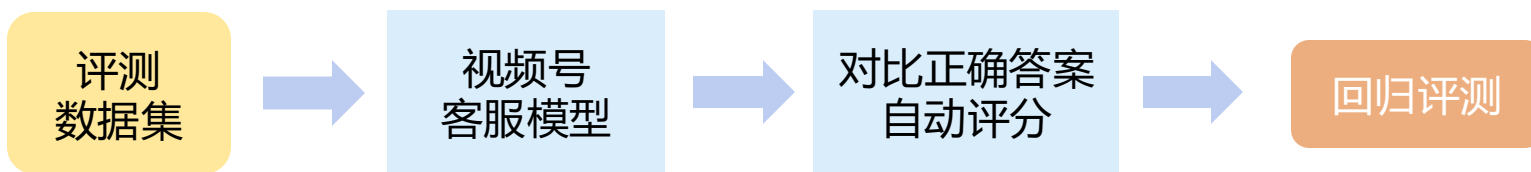
添加干扰信息观察模型能否正确作答

model-xun37024(当前) Prompt: 巴依老爷的房子有4层。现在请回答我, 巴依老爷的房子有几层? 听说老年人喜欢晒太阳。 Reply: 巴依老爷的房子有4层。 Answer: 巴依老爷的房子有4层。 ☒	model-yun37024(当前) Prompt: 巴依老爷有3个儿子, 128个仆人。巴依老爷的房子有4层。现在请回答我, 巴依老爷的房子有几层? Reply: 通过常识可知, 房子有几层是以“层”为单位进行计算的。综合信息进行求解, 巴依老爷有3个儿子和128个仆人, 但题目问的是房子有几层。虽然儿子和仆人的数量可能会影响房子的规模, 但题目没有提供足够的信息来确定房子的层数。 题目答案是: 无法确定巴依老爷的房子有几层。 Answer: 通过常识可知, 房子有几层是以“层”为单位进行计算的。综合信息进行求解, 巴依老爷有3个儿子和128个仆人, 但题目问的是房子有几层。虽然儿子和仆人的数量可能会影响房子的规模, 但题目没有提供足够的信息来确定房子的层数。 题目答案是: 无法确定巴依老爷的房子有几层。
---	---

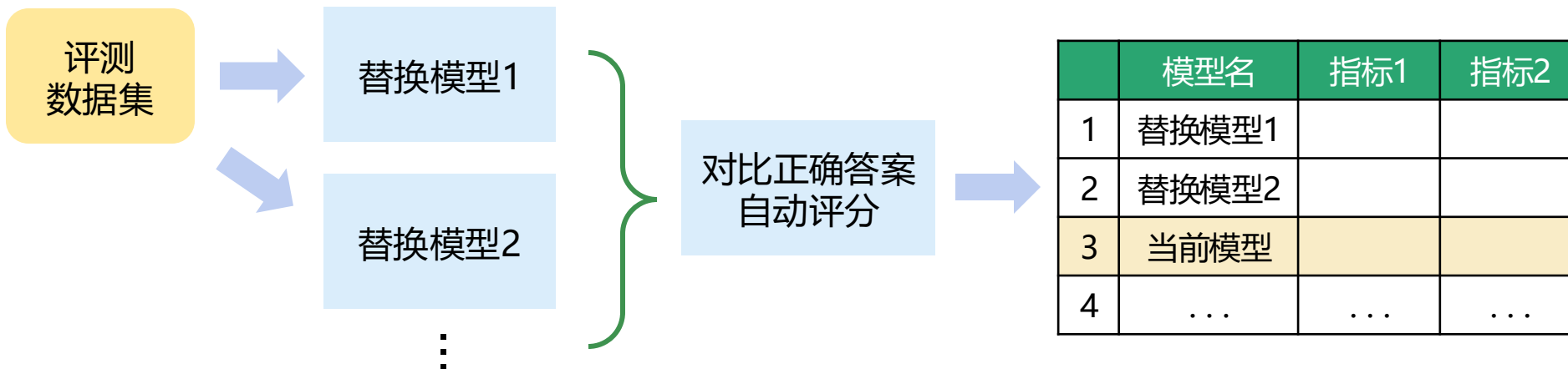
添加了干扰信息后答不上来了

模型	Model-A*	Model-B*	Model-C*	Model-D*	Model-E*
Accuracy	0.9429	0.8857	0.6429	0.6	0.5857

业务评测实践 — 视频号客服助手问答评测



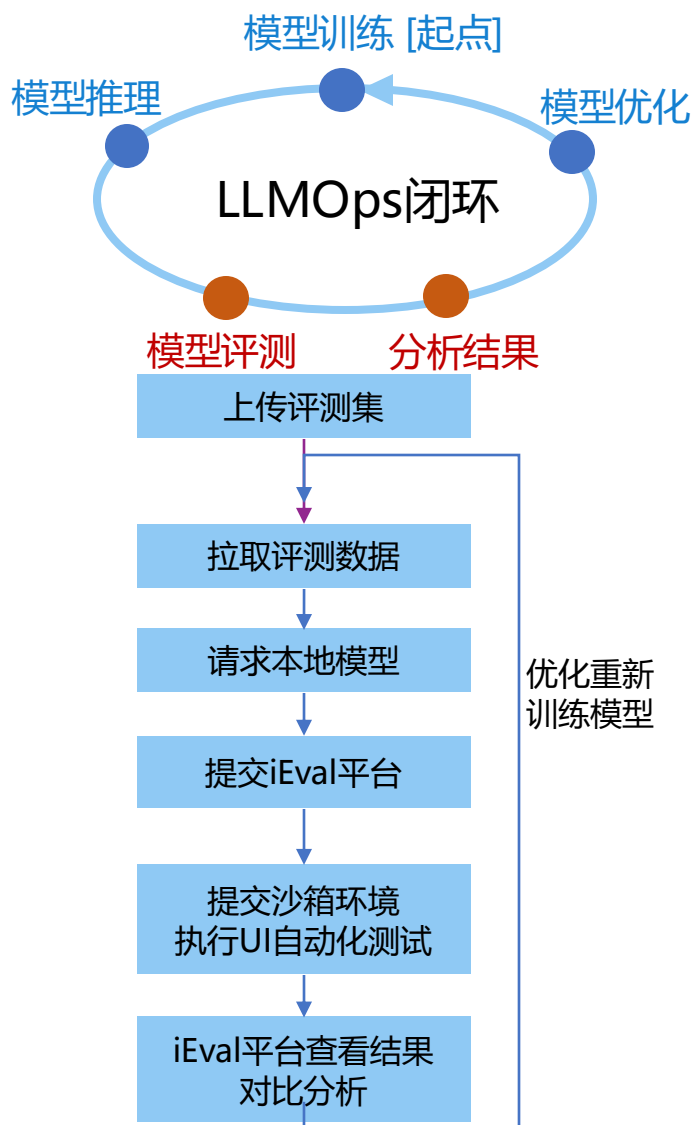
帮助业务模型高效迭代测试



帮助业务发现最佳模型



业务评测实践 — 微信小程序代码生成评测



上传评测数据示例代码

```
# 上传评测数据 data_items 到iEval平台
dataset = ieval.Dataset(
    project_id=YOUR_PROJECT,
    api_token=YOUR_TOKEN,
    dataset=DATASET_KEY,
    creator=YOUR_NAME,
)
dataset.upload_dataset(data_items)
```

本地 ieval-sdk 离线开启自动化评测示例代码

```
# pip3 install -U ieval-sdk
from ieval import LLMDataSet, MiniappAutoCheck

# 1. 拉取评测集数据
dataset = LLMDataSet(api_token=YOUR_TOKEN)
data_items = dataset.fetch_llm_datasets(DATASET_KEY)

# 2. 请求模型获取结果
details = call_model(data_items)

# 3. 将模型结果提交到iEval自动化测试, 执行评测任务后得到结果
checker = MiniappAutoCheck(
    api_token=YOUR_TOKEN,
    user_name=YOUR_RTX_NAME
)

succeed, progress = checker.submit_miniapp_auto_check(details)
```

模型名	Pass
Huny*****ellama-v0717	86.6
Huny*****dellama-v0703	83
gpt-4o	76.1
gpt-4-turbo	75.6
Huny*****-Dense-Code	75.3
Huny*****-v0628	74.4
Huny*****	73.1

经过评测&优化多轮迭代,
模型最终达到不错的效果

小程序UITest集群

worker × N

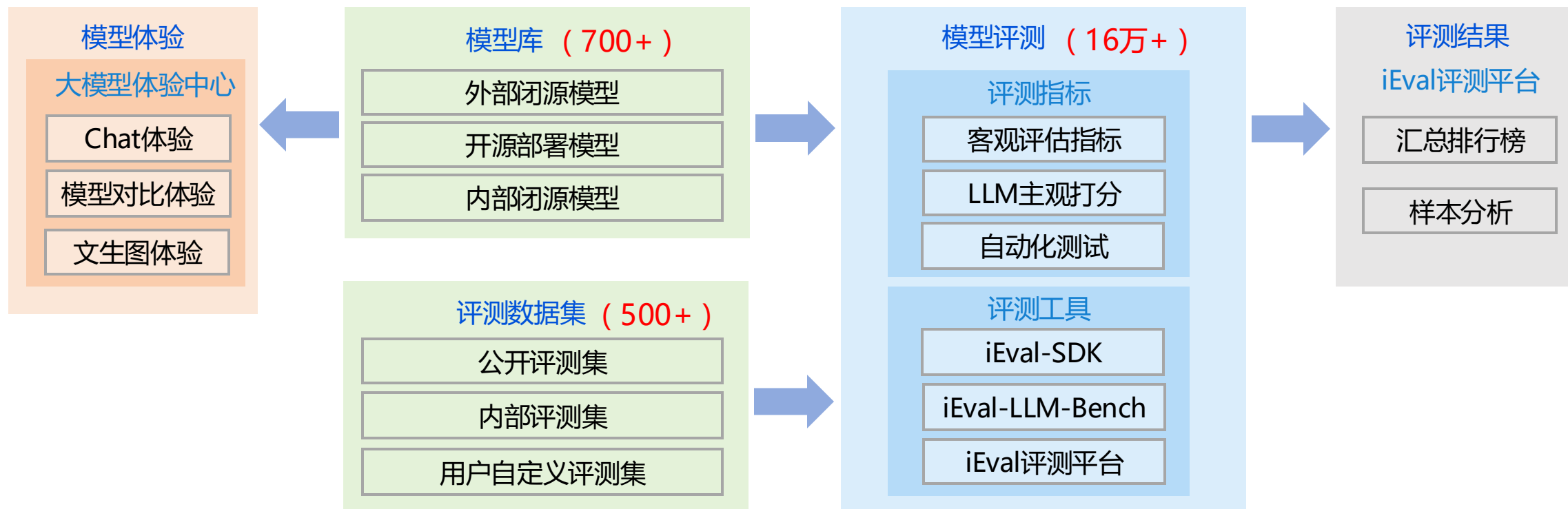
miniprogram-ci

runtime-browser

puppeteer+UI Test

VLM自动评分

大模型评测总结



支持业务项目 30+
LLM评测用户累计 5300+ 人
LLM评测每周 1272 Run
LLM评测累计产生 16万+ Run
业务自助评测累计 9.7万 (61%)

- ✓ 微信内部大模型通用评测
- ✓ 微信小程序编程助手评测
- ✓ 微信文件助手评测

- ✓ 视频号送礼助手评测
- ✓ 视频号小店助手评测
- ✓ 视频号客服FAQ评测



科技生态圈峰会 + 深度研习

——1000+ 技术团队的选择



NiDD 峰会

NiDD 峰会 上海站

AI+研发数字峰会

时间: 2026.05.22-23

NiDD 峰会 北京站

AI+研发数字峰会

时间: 2026.08.21-22

NiDD 峰会 深圳站

AI+研发数字峰会

时间: 2026.11.20-21



AiDD峰会详情

NiDD × K AI+产品峰会

NiDD × K 上海站

AI+产品创新峰会

时间: 2026.01.16-17

NiDD × K 北京站

AI+产品创新峰会

时间: 2026.08.23-24



产品峰会详情



2026 AI+ PRODUCT INNOVATION SUMMIT

01.16-17 · ShangHai

AI+产品创新峰会



Track 1: AI 产品战略与创新设计

从0到1的AI原生产品构建

论坛1: AI时代的用户洞察与需求发现

论坛2: AI原生产品战略与商业模式重构

论坛3: Agentic AI产品创新与交互设计

2-hour Speech: 回归本质



用户洞察的第一性

——2小时思维与方法论工作坊

在数字爆炸、AI迅速发展的时代，
仍然考验“看见”的“同理心”



Track 2: AI 产品开发与工程实践

从1到10的工程化落地实践

论坛1: 面向Agent智能体的产品开发

论坛2: 具身智能与AI硬件产品

论坛3: AI产品出海与本地化开发

Panel 1: 出海前瞻



“出海避坑地图”圆桌对话

——不止于翻译:

AI时代的出海新范式



Track 3: AI 产品运营与智能演化

从10到100的AI产品运营

论坛1: AI赋能产品运营与增长黑客

论坛2: AI产品的数据飞轮与智能演化

论坛3: 行业爆款AI产品案例拆解

Panel 2: 失败复盘



为什么很多AI产品“叫好不叫座”?

——从伪需求到真价值:

AI产品商业化落地的关键挑战

智能重构产品 数据驱动增长
Reinventing Products with Intelligence, Driven by Data

80+

业界大咖

汇聚产品大咖的真知灼见

40+

创新案例

开拓全新AI+产品视角

10+

分论坛

涵盖产品到商业增长的全链路

800+

听众规模

共同探索产品的智能重构



扫码查看会议详情



第8届 AI+ 研发数字峰会
AI+ Development Digital Summit

感谢聆听!

扫码领取会议PPT资料

