

基础岛任务笔记1-书生大模型全链路开源体系

1 任务

- 2 观看本关卡视频和官网<https://internlm.intern-ai.org.cn/>后，写一篇关于书生大模型全链路开源开放体系的笔记发布到知乎、CSDN等任一社交媒体，将作业链接提交到以下问卷，助教老师批改后将获得 100 算力点奖励！！！
- 3 提交地址：<https://aicarrier.feishu.cn/share/base/form/shrcnZ4bQ4YmhEtMtnKxZUcf1vd>

前言

InternLM也就是**书生·浦语大模型**，上海人工智能实验室开发的大语言模型。通过创新的预训练和优化技术，在6个维度和30个基准的综合评估、长期上下文建模和开放式主观评估方面超过了ChatGPT。

经过去年七月份到今年一年的努力，打通了**大模型全链路体系**：

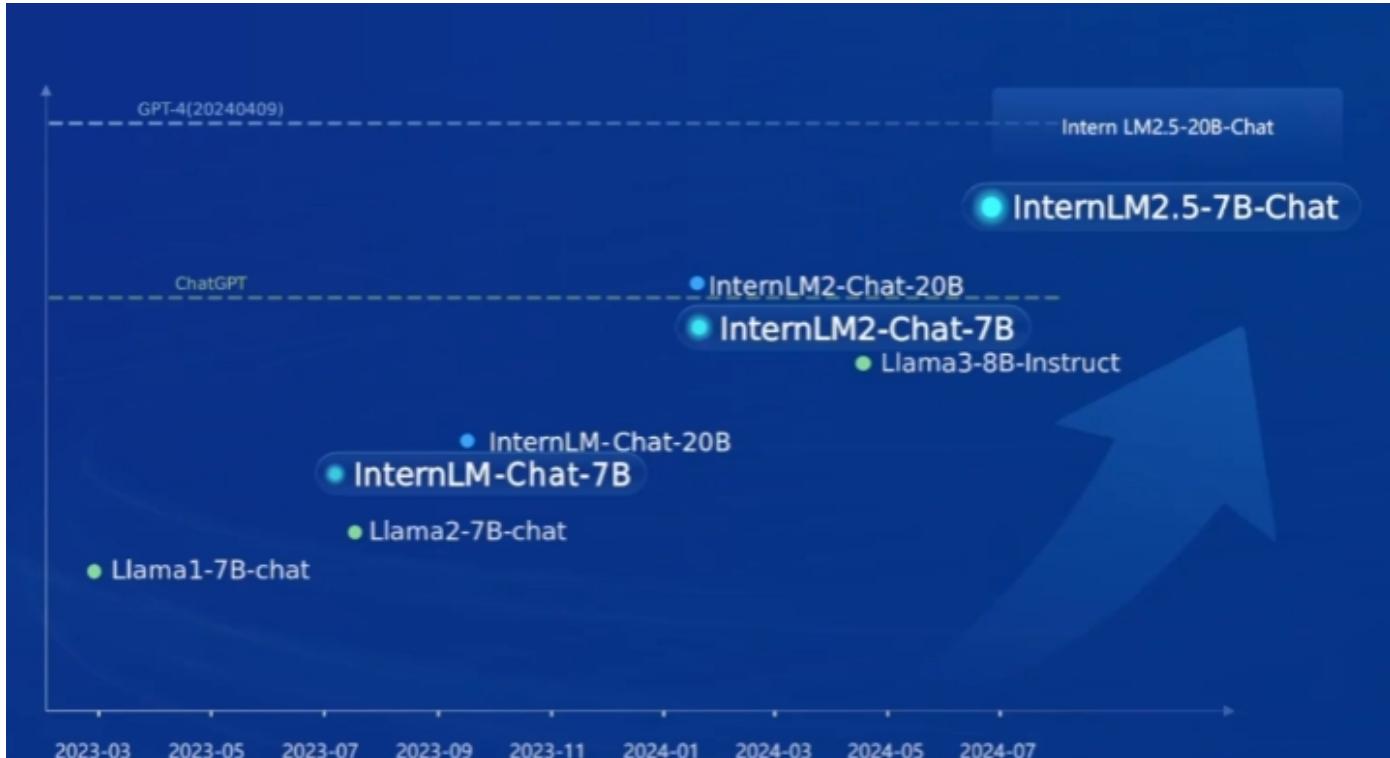
数据的收集整理->数据标注->模型训练->模型微调->模型评测->基于模型的agent rag 搜索引擎

InternLM简述

- 开源历程：



- 性能天梯图(与ChatGPT进行比较)：



InternLM2.5概览

- 相对于版本2.0性能提升20%，支持100万字上下文，自主规划和搜索完成复杂任务效率提升60倍。
- 开源体系迭代过程核心思想
 - 流程



- 数据构造（开源数据标注项目 open data lab）

高质量合成数据

基于规则的数据构造

基于模型的数据扩充

基于反馈的数据生成

由于 $\sin A = \sin(\frac{5\pi}{12})$,
我们可以使用和角公式来找到 $\sin A$:

$$\sin A = \sin(\frac{5\pi}{12}) = \sin(\frac{\pi}{4} + \frac{\pi}{6}) = \sin \frac{\pi}{4} \cos \frac{\pi}{6} + \cos \frac{\pi}{4} \sin \frac{\pi}{6} = \frac{\sqrt{2}}{2} + \frac{\sqrt{3}}{2} + \frac{\sqrt{2}}{2} + \frac{1}{2} = \frac{\sqrt{6} + \sqrt{2}}{4}$$

床前明月光，疑是地上霜。举头望明月，低头思故乡。
没有创作
平仄错误
符合要求

融合多种数据合成方案，提升合成数据质量

- 2.5版本7B模型解决问题

- 推理能力提升



- 模型解决问题思路（在跨文档方面超过RAG）



- 基于规划和搜索解决复杂问题（开源了调用搜索引擎的工具）
- 书生·浦语开源模型谱系



社区生态



• 数据

飞速成长

- 模态** 30+
- 数据集** 7,700+
- 数据大小** 180TB

丰富多样的开放数据

- 60亿 图像** LAION-5B SA-1B ImageNet
- 1万亿 tokens语料** The Pile C4 WikiQA
- 2万 小时音频** LibriSpeech VoxCeleb Speech Commands
- 8亿 片段视频** MovieNet Kinetics MOT
- 1百万 3D 模型** OmniObject3D ShapeNet Scannet

服务与工具

- 灵活检索** 支持 10+ 搜索条件组合
- 高速下载** 单文件稳定速度至少 20M/s
- 智能标注** 支持 30+ 工具组合形式
- 高效采集** 整体效率可提升 40%

• 开源数据处理工具箱

Miner U

一站式开源高质量数据提取工具，支持多格式（PDF/网页/电子书），智能萃取，生成高质量预训练/微调语料。

- 复杂版面/公式精准识别
- 性能超过商业软件

Miner U PDF 文档提取工具

<https://github.com/opendatalab/MinerU>

Label LLM

专业致力于 LLM 对话标注，通过灵活多变的工具配置与多种数据模态的广泛兼容，为大模型量身打造高质量的标注数据。

- 支持指令采集、偏好收集、对话评估……
- 多人协作、任务管理、源码开放可魔改

Label LLM 智能对话标注

<https://github.com/opendatalab/LabelLLM>

Label U

一款轻量级开源标注工具，自由组合多样工具，无缝兼容多格式数据，同时支持载入预标注，加速数据标注效率。

- 支持图片、视频、音频多种数据标注
- 小巧灵活，AI 标注导入二次人工精修

Label U 智能标注

<https://github.com/opendatalab/labelU>

- 预训练框架InternEvo



- 微调框架 XTuner



- 大模型评测体系 OpenCompass 司南

广泛应用于头部大模型企业和科研机构



大模型评测国标 主要参与单位



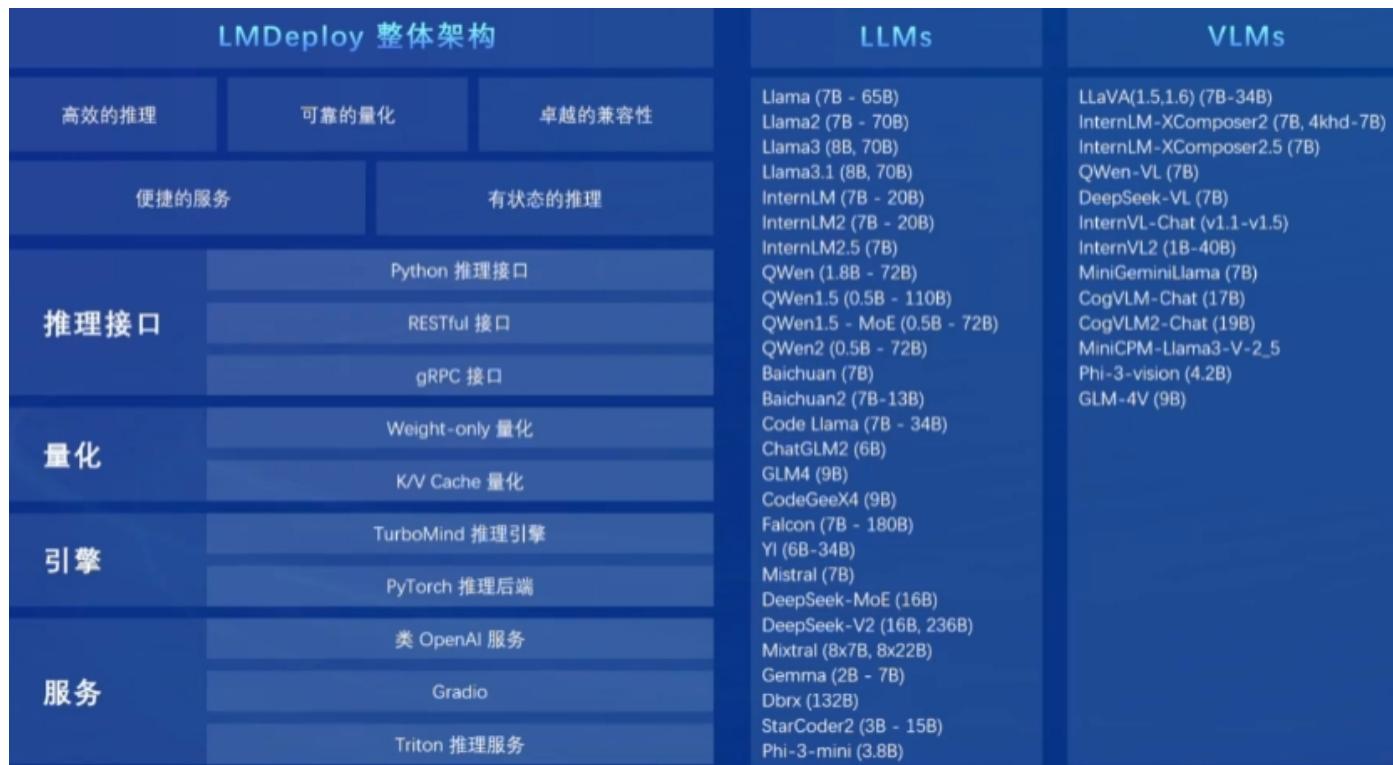
获得 Meta 官方推荐 唯一国产大模型评测体系

These types of projects provide a quantitative way of looking at the models performance in simulated real world examples. Some of these projects include the [LM Evaluation Metrics](#) (used to create the [leaderboard](#)), [HFLM](#), [Bio-BERT](#) and [OpenCompass](#).

开源社区最完善的评测体系之一 超过100+ 评测集 50万+ 题目



• 大模型部署框架LMDeploy

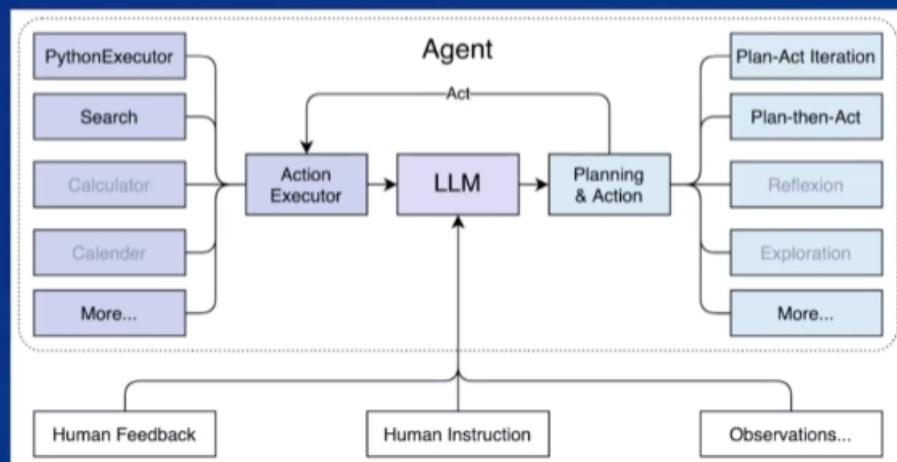


• 智能体框架 Lagent

大语言模型的局限性

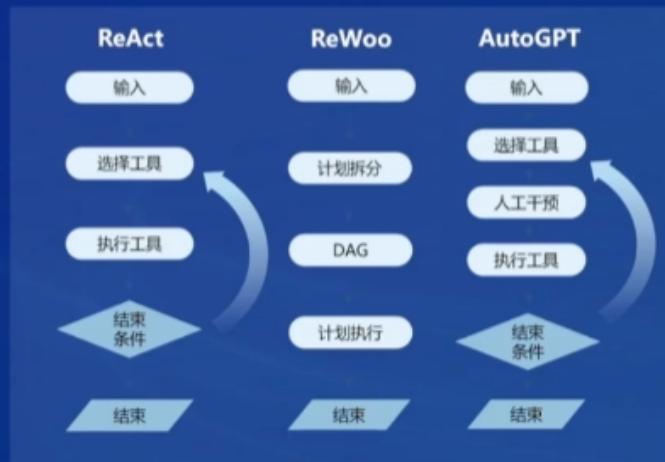
LLM → 智能体

- 最新信息和知识的获取
- 回复的可靠性
- 数学计算
- 工具使用和交互



轻量级智能体框架 Lagent

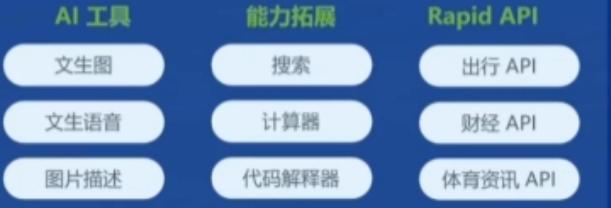
支持多种类型的智能体能力



灵活支持多种大语言模型



简单易拓展，支持丰富的工具



- 最新开源的智能体MindSearch（搜索引擎）
- 企业级知识库构建工具HuixiangDou

介绍

HuixiangDou 是群聊场景 LLM 知识助手，为即时通讯（如微信、飞书）群聊场景设计。



检索增强生成 (RAG)，非参数记忆，利用外部知识库提供实时更新的信息。



结构化知识库匹配意图
行为可解释。

场景特点



无关问题不吭声



明确回答的直接回复



不违背核心价值观

特性



开源
BSD-3-Clause
免费商用



实战派
应用 RAG 和 KG
1500+ 知识库
500+ 用户群
业务数据实测精度



领域知识
7 种文档格式
更新立即生效



安全
支持私有化部署
数据不上传



简单便宜
最低仅 2G 显存
支持现有客户群



扩展性强
2 类 IM 软件
9 个 LLM 接口