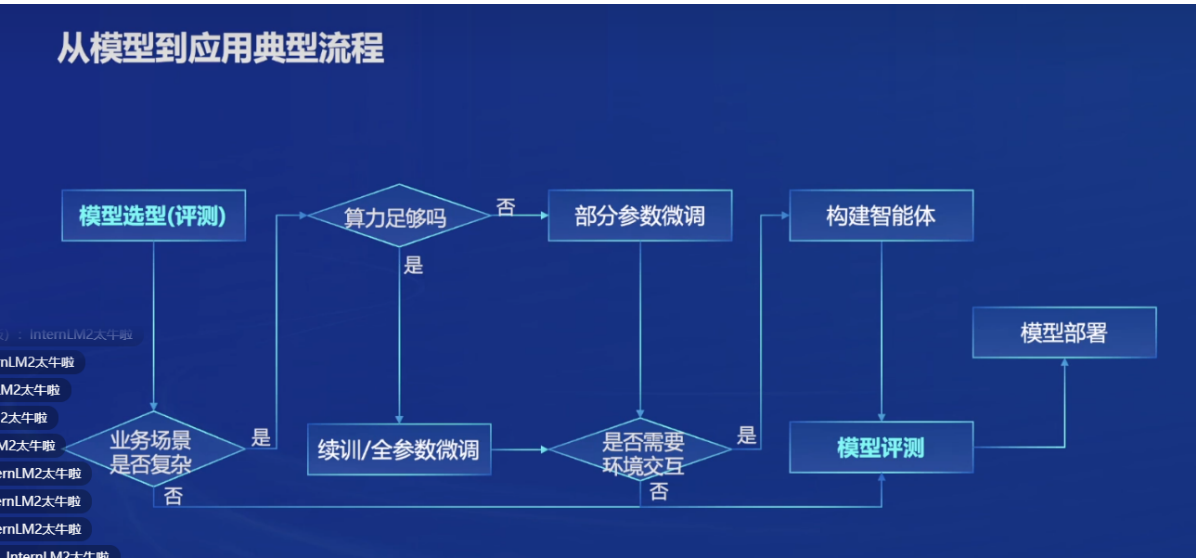


前面主要介绍了InterLM2的一些应用场景相比于上一代的提升



全链条开源：



开源数据集：



Xtuner介绍：

全链条开源开放体系 | 微调

高效微调框架 XTuner

InternLM

Llama

Qwen

BaiChuan

ChatGLM

任务类型

增量预训练

指令微调

工具类指令微调

数据格式

Alpaca

MOSS

OpenAI

Guanaco

...

训练引擎

ML Engine

优化加速

Flash Attention

DeepSpeed ZeRO

Pytorch FSDP

支持算法

QLoRA 微调

LoRA 微调

全量参数微调

消费级显卡

GeForce RTX 2080、2080ti

GeForce RTX 3060 ~ 3090ti

GeForce RTX 4060 ~ 4090

数据中心

Tesla T4、V100

A10、A100、H100

适配多种生态

• 多种微调算法

多种微调策略与算法，覆盖各类 SFT 场景

• 适配多种开源生态

支持加载 HuggingFace、ModelScope 模型或数据集

• 自动优化加速

开发者无需关注复杂的显存优化与计算加速细节

适配多种硬件

• 训练方案覆盖 NVIDIA 20 系以上所有显卡

• 最低只需 8GB 显存即可微调 7B 模型

评测榜单：

CompassRank：中立全面的性能榜单

OpenCompass

数据集社区

评测榜单

评测工具

文档

中 | EN

加入评测

贡献数据集

登录

CompassRank 将未来的可能性定位在今天

致力于探索最先进的语言与视觉模型，为工业界和研究社区提供全面、客观、中立的评测参考

大语言模型总榜

全部

24-01

23-12

多模态模型总榜

全部

24-01

23-12

NLP

大语言模型评测

100+ 大语言模型已加入评测

包含开源模型与闭源模型（如 GPT 系列模型），涵盖从 0.1B 到 175B 参数的语言模型

模型	模型
OpenAI	GPT-4o / GPT-4o / GPT-3.5-Turbo
阿里集团	通义千问
智谱 AI	ChatGLM
Meta	LLaMA
上海人工智能实验室	书生·浦语
百川智能	BaiChuan
零一万物	Yi
加州大学圣迭戈分校	Vicuna

查看全部评测榜单 >>>

\_NLP

多模态模型评测

50+ 多模态大模型已加入评测

包含开源模型与闭源模型（如 GPT 系列模型），涵盖从 0.1B 到 175B 参数的多模态模型

模型	模型
Google	GeminiProVision
OpenAI	GPT-4o
阿里集团	Qwen-VL / Qwen-VL-Chat / Qwen-VL
上海人工智能实验室	InternLM-XTuner-VL / LLaVA-InternLM / (70B/130B) / (70B/130B)
智谱 AI	CogVLM-FTS-Chat / VisualGLM-400
华中科技大学	Monkey / Monkey-Chat
微软亚洲研究院	LLaVA-v1.5-70B/130B
HuggingFace	LLaVA-1.5-70B/130B

查看全部评测榜单 >>>

OpenCompass:

## OpenCompass 助力大模型产业发展和学术研究

### 广泛应用于头部大模型企业和科研机构

Alibaba

HUAWEI

招商银行

中国平安

AI2

BOSS直聘

小红书

Baidu 百度

Microsoft

中国电信

复旦大学

oppo

东方财富

美团

Tencent 腾讯

vivo

百川智能

MINIMAX

NANYANG TECHNOLOGICAL UNIVERSITY

#### 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 Harness](#) (used to create the [HF leaderboard](#)), [HELM](#), [BIG-bench](#) and [OpenCompass](#).

#### 社区支持最完善的评测体系之一

#### 100+ 评测集 50万+ 题目

学科

语言

知识

理解

推理

理科能力，性能与尺寸呈现强相关性

### 洞见未来：OpenCompass 年度榜单(综合性客观评测)

The chart displays the performance of various large language models (LLMs) across five categories: Language (语言), Knowledge (知识), Reasoning (推理), Mathematics (数学), and Code (代码). The models are ranked from highest to lowest overall score. GPT-4 Turbo is the top performer, followed by Qwen2.5-72B-Instruct and Qwen2.5-72B-Instruct-Chat. The chart shows that while some models perform well in language and knowledge tasks, they struggle in reasoning and mathematics. The overall trend shows that as the model size increases, the performance generally improves, especially in reasoning and mathematics tasks.

Model	语言	知识	推理	数学	代码
GPT-4 Turbo	65	60	55	50	45
Qwen2.5-72B-Instruct	60	55	50	45	40
Qwen2.5-72B-Instruct-Chat	55	50	45	40	35
Qwen2.5-72B-Instruct-Chat-2	50	45	40	35	30
Qwen2.5-72B-Instruct-Chat-3	45	40	35	30	25
Qwen2.5-72B-Instruct-Chat-4	40	35	30	25	20
Qwen2.5-72B-Instruct-Chat-5	35	30	25	20	15
Qwen2.5-72B-Instruct-Chat-6	30	25	20	15	10
Qwen2.5-72B-Instruct-Chat-7	25	20	15	10	5
Qwen2.5-72B-Instruct-Chat-8	20	15	10	5	0
Qwen2.5-72B-Instruct-Chat-9	15	10	5	0	0
Qwen2.5-72B-Instruct-Chat-10	10	5	0	0	0

#### 整体能力仍有较大提升空间

采用了更加准确的循环评测策略，我们实现了对模型真实能力分析，在百分制的客观评测基准中，GPT-4-Turbo也仅仅达到了61.8分的及格水平。

#### “理科”能力和模型尺寸关联性高

在语言和知识这类“文科”维度，中轻量级模型和重量级/闭源商业模型差距较小，但数学、推理、代码等维度上，性能和尺寸呈现较强相关性

#### 复杂推理仍是短板

国内多个模型综合能力和GPT-4-Turbo在接近，但在复杂推理上仍然存在较大差距，并且和模型尺寸存在较强相关性。

#### 模型主客观性能需综合参考

大量开源模型和API模型的客观性能和主观性能存在较大的偏差，社区不仅仅需要夯实客观能力基础，更需要在偏好对齐和对话体验上下功夫。

## 全链条开源开放体系 | 智能体

### 多模态智能体工具箱 AgentLego

- 丰富的工具集合，尤其是提供了大量视觉、多模态相关领域的前沿算法功能
- 支持多个主流智能体系统，如 LangChain, Transformers Agent, lagent 等
- 灵活的多模态工具调用接口，可以轻松支持各类输入输出格式的工具函数
- 一键式远程工具部署，轻松使用和调试大模型智能体

