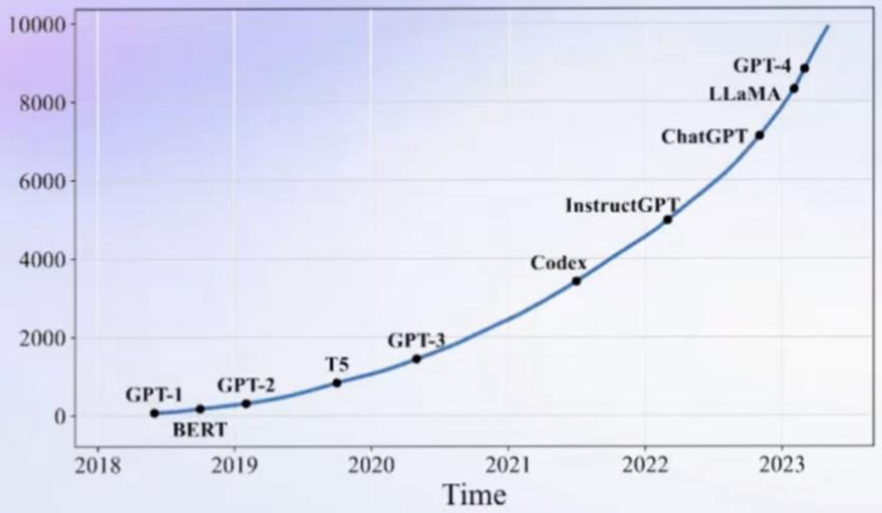
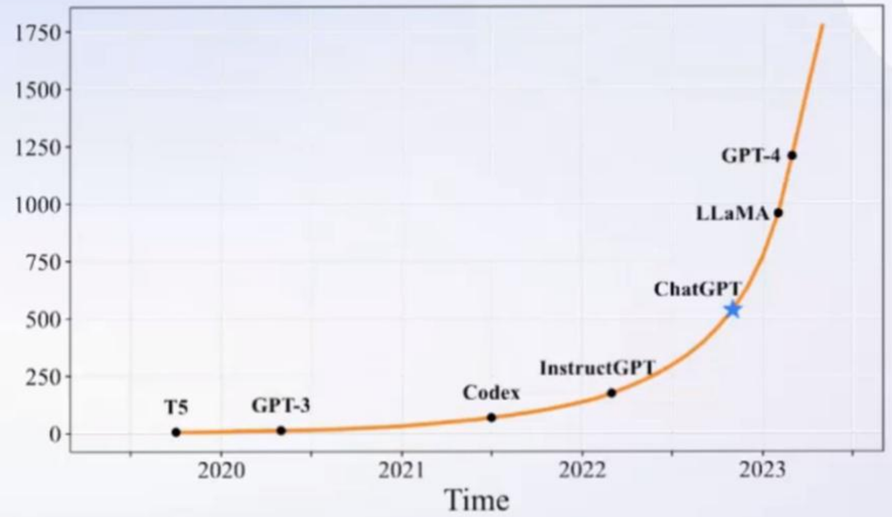
书生·浦语大模型全链路开源体系

陈恺 | 上海人工智能实验室 青年科学家

大模型成为热门关键词

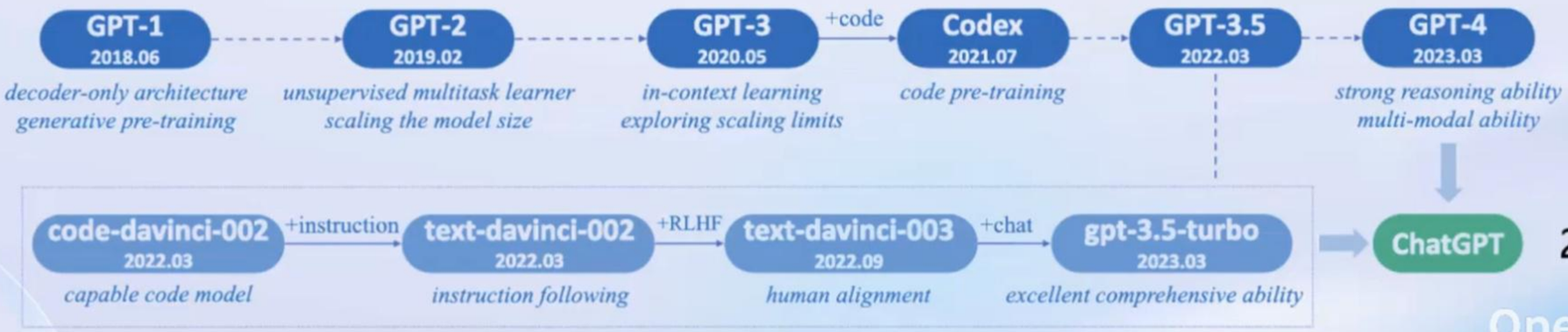


(a) Query="Language Model"



(b) Query="Large Language Model"

OpenAI



2个月月活破亿

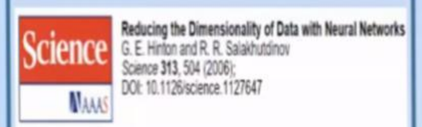
Reference: A Survey of Large Language Models, Zhao et al.

大模型成为发展通用人工智能的重要途径

专用模型：
针对特定任务，一个模型解决一个问题

通用大模型：
一个模型应对多种任务、多种模态

2006 深度学习理论突破



深度置信网络

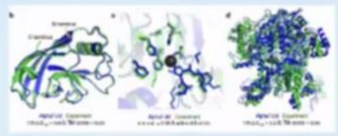
2012 ImageNet竞赛

1000类，100万数据

2016 围棋比赛

AlphaGo 4:1 李世乭

2021 AlphaFold



蛋白质结构预测准确率新高

2011 大规模语音识别

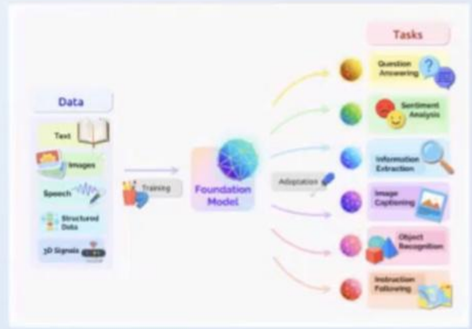
Switchboard错误降低9%

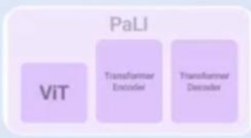
2014 人脸识别

LFW识别率99%,超过人类

2019 德州扑克

首次在多人复杂对局中超越人类





ChatGPT GPT-4

书生·浦语大模型开源历程

6月7日

InternLM

千亿参数语言大模型发布

8月14日

书生·万卷 1.0

多模态预训练语料库开源发布

8月28日

InternLM 千亿参数模型

参数量升级至 123B

7月6日

InternLM

千亿参数大模型全面升级

支持8K语境、26种语言

全面开源，免费商用：

InternLM-7B 模型、全链条开源工具体系

8月21日

升级版对话

模型 InternLM-Chat-7B v1.1发布

开源智能体框架**Lagent**

支持从语言模型到智能体升级转换

9月20日

增强版InternLM-20B开源

开源工具链全线升级

书生·浦语大模型系列

轻量级：InternLM-7B

- 70亿模型参数，小巧轻便，便于部署
- 10000亿训练token数据，信息全面，能力多维
- 具备长语境能力，支持8k语境窗口长度
- 具备通用工具调用能力，支持多种工具调用模板

中量级：InternLM-20B

- 200亿参数量，在模型能力与推理代价间取得平衡
- 采用深而窄的结构，降低推理计算量但提高了推理能力
- 4k训练语境长度，推理时可外推至16k

重量级：InternLM-123B

- 1230亿模型参数，强大的性能
- 具备极强的推理能力、全面的知识覆盖面、超强理解能力与对话能力
- 准确的API调用能力，可实现各类Agent

社区低成本可用最佳模型规模

商业场景可开发定制高精度较小模型规模

通用大语言模型能力全面覆盖千亿模型规模

书生·浦语20B开源大模型性能

能力	测试集	Llama-13B	Llama2-13B	Baichuan2-13B	**InternLM-20B**	Llama-33B	Llama-65B	Llama2-70B
综合考试	MMLU	47.73	54.99	59.55	**62.05**	58.73	63.71	69.75
	C-Eval (val)	31.83	41.4	**59.01**	58.8	37.47	40.36	50.13
	AGI-Eval	22.03	30.93	37.37	**44.58**	33.53	33.92	40.02
知识问答	BoolQ	78.75	82.42	67	**87.46**	84.43	86.61	87.74
	TriviaQA	52.47	59.36	46.61	57.26	**66.24**	69.79	70.71
	NaturalQuestions	20.17	24.85	16.32	25.15	**30.89**	33.41	34.16
阅读理解	CMRC	9.26	31.59	29.85	**68.78**	14.17	34.73	43.74
	CSL	55	58.75	63.12	**65.62**	57.5	59.38	60
	RACE (middle)	53.41	63.02	68.94	**86.35**	64.55	72.35	81.55
	RACE (high)	47.63	58.86	67.18	**83.28**	62.61	68.01	79.93
	XSum	20.37	23.37	25.23	**35.54**	20.55	19.91	25.38
推理	WinoGrande	64.64	64.01	67.32	**69.38**	66.85	69.38	69.77
	BBH	37.93	45.62	48.98	**52.51**	49.98	58.38	64.91
	GSM8K	20.32	29.57	**52.62**	**52.62**	42.3	54.44	63.31
	PIQA	79.71	79.76	78.07	80.25	**81.34**	82.15	82.54
编程	HumanEval	14.02	18.9	17.07	**25.61**	17.68	18.9	26.22
	MBPP	20.6	26.8	30.8	**35.6**	28.4	33.6	39.6

- 全面领先相近量级的开源模型（包括Llama-33B、Llama2-13B以及国内主流的7B、13B开源模型）
- 以不足三分之一的参数量，达到Llama2-70B水平

从模型到应用

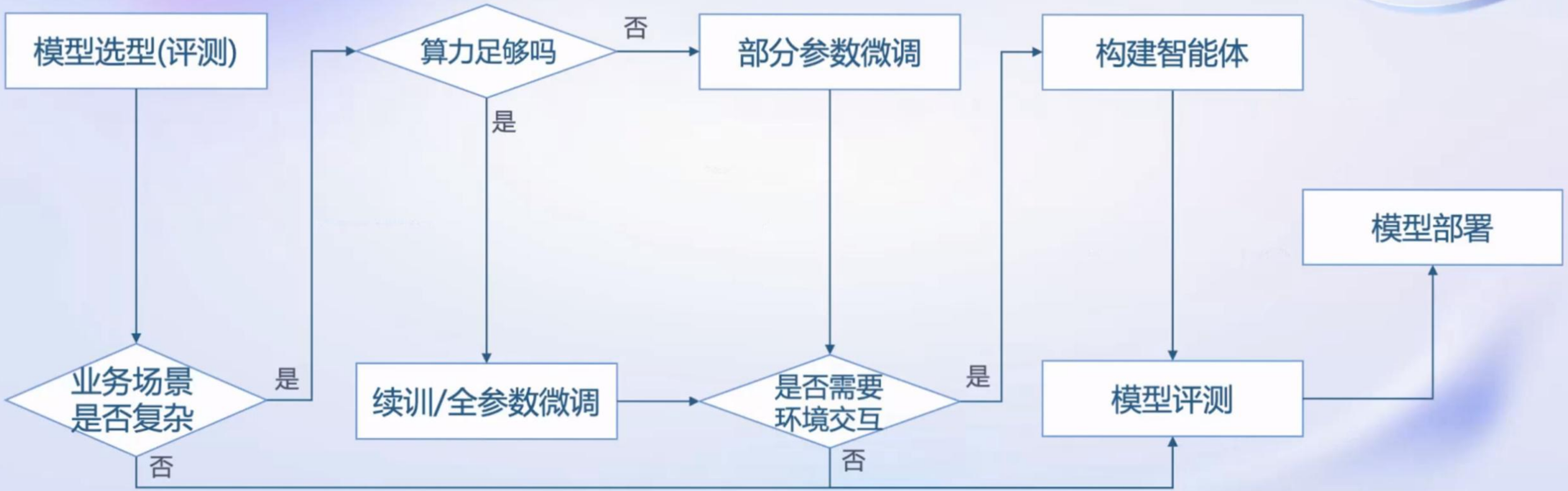
书生·浦语

智能客服

个人助手

行业应用

从模型到应用


模型选型(评测)
算力足够吗
否
部分参数微调
构建智能体
模型部署
是
业务场景
是否复杂
是
续训/全参数微调
是否需要
环境交互
是
模型评测
否
否

书生·浦语全链条开源开放体系

数据

书生·万卷

2TB数据，
涵盖多种模态与任务

预训练

InternLM-Train

并行训练，极致优化
速度达到 3600 tokens/sec/gpu

微调

XTuner

支持 全参数微调，
支持LoRA等低成本微调

部署

LMDeploy

全链路部署，性能领先
每秒生成 2000+ tokens

评测

OpenCompass

全方位评测，性能可复现
80 套评测集，40 万道题目

应用

Lagent
AgentLego

支持多种智能体，支持代码解释器等多种工具

全链条开源开放体系 | 数据

书生· 万卷 1.0

文本数据

- 50亿个文档
- 数据量超 1TB

图像-文本数据集

- 超2,200万个文件
- 数据量超140GB

视频数据

- 超1,000个文件
- 数据量超900GB

总数据量：2TB

发布日期：8月14日

多模态融合

万卷包含文本、图像和视频等多模态数据，涵盖科技、文学、媒体、教育和法律等多个领域。该数据集对模型的知识内容、逻辑推理和泛化能力的提升有显著效果。

精细化处理

万卷经过语言筛选、文本提取、格式标准化、数据过滤和清洗（基于规则和模型）、多尺度去重和数据质量评估等精细数据处理环节，能够很好地适应后续模型训练的要求。

价值观对齐

在万卷的构建过程中，研究人员注重将数据内容与主流中国价值观进行对齐，并通过算法和人工评估的结合提高语料库的纯净度。

全链条开源开放体系 | 数据

飞速成长

模态	30+
数据集	5,400+
数据大小	80TB

OpenDataLab 丰富多样的开放数据

60 亿 图像
LAION-5B SA-1B ImageNet

8 亿 片段 视频
MovieNet Kinetics MOT

1 万亿 tokens 语料
The Pile C4 WikiQA

1 百万 3D 模型
OmniObject3D ShapeNet Scannet

2 万 小时 音频
LibriSpeech VoxCeleb Speech Commands

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%

全链条开源开放体系 | 预训练

高可扩展

支持从 8 卡到千卡训练 千卡加速效率达 92%

极致性能优化

Hybrid Zero 独特技术 + 极致优化，加速 50%

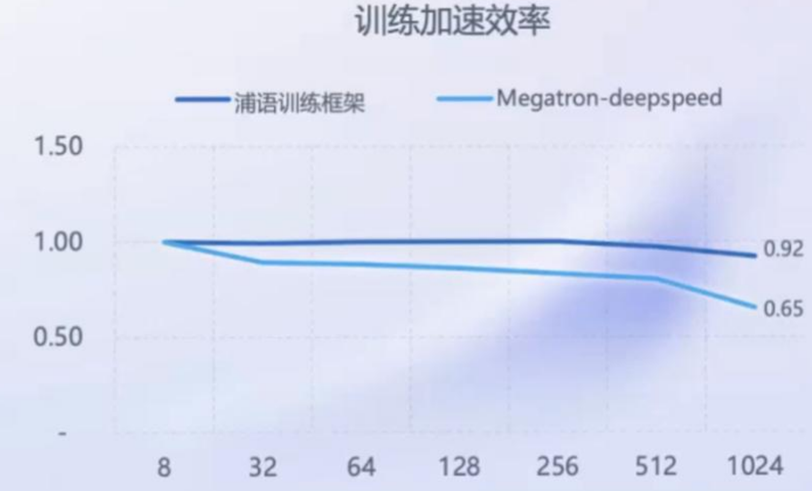
兼容主流

无缝接入 HuggingFace 等技术生态，支持各类轻量化技术

开箱即用

支持多种规格语言模型，修改配置即可训练

训练算法	预训练	微调
训练优势	高性能 Transformer 计算库	多种并行策略
通信/计算调度	梯度累积 算法选择	通信/计算重叠
显存管理	优化器状态 梯度	参数



全链条开源开放体系 | 微调

大语言模型的下游应用中，增量续训和有监督微调是经常会用到两种方式。

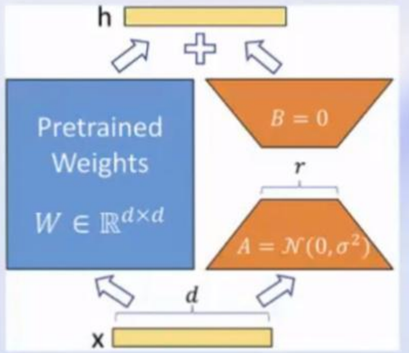
增量续训

使用场景：让基座模型学习到一些新知识，如某个垂类领域知识
训练数据：文章、书籍、代码等

全量参数微调　　部分参数微调

有监督微调

使用场景：让模型学会理解和遵循各种指令，或者注入少量领域知识
训练数据：高质量的对话、问答数据



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	数据格式	训练引擎	优化加速	支持算法
增量预训练 指令微调 工具类指令微调	Alpaca MOSS OpenAI Guanacao ...	MMEngine	Flash Attention DeepSpeed ZeRO Pytorch FSDP	QLoRA 微调 LoRA 微调 全量参数微调

NVIDIA

消费级显卡	数据中心
GeForce RTX 2080、2080ti	Tesla T4、V100
GeForce RTX 3060 ~ 3090ti	A10、A100、H100
GeForce RTX 4060 ~ 4090	

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**

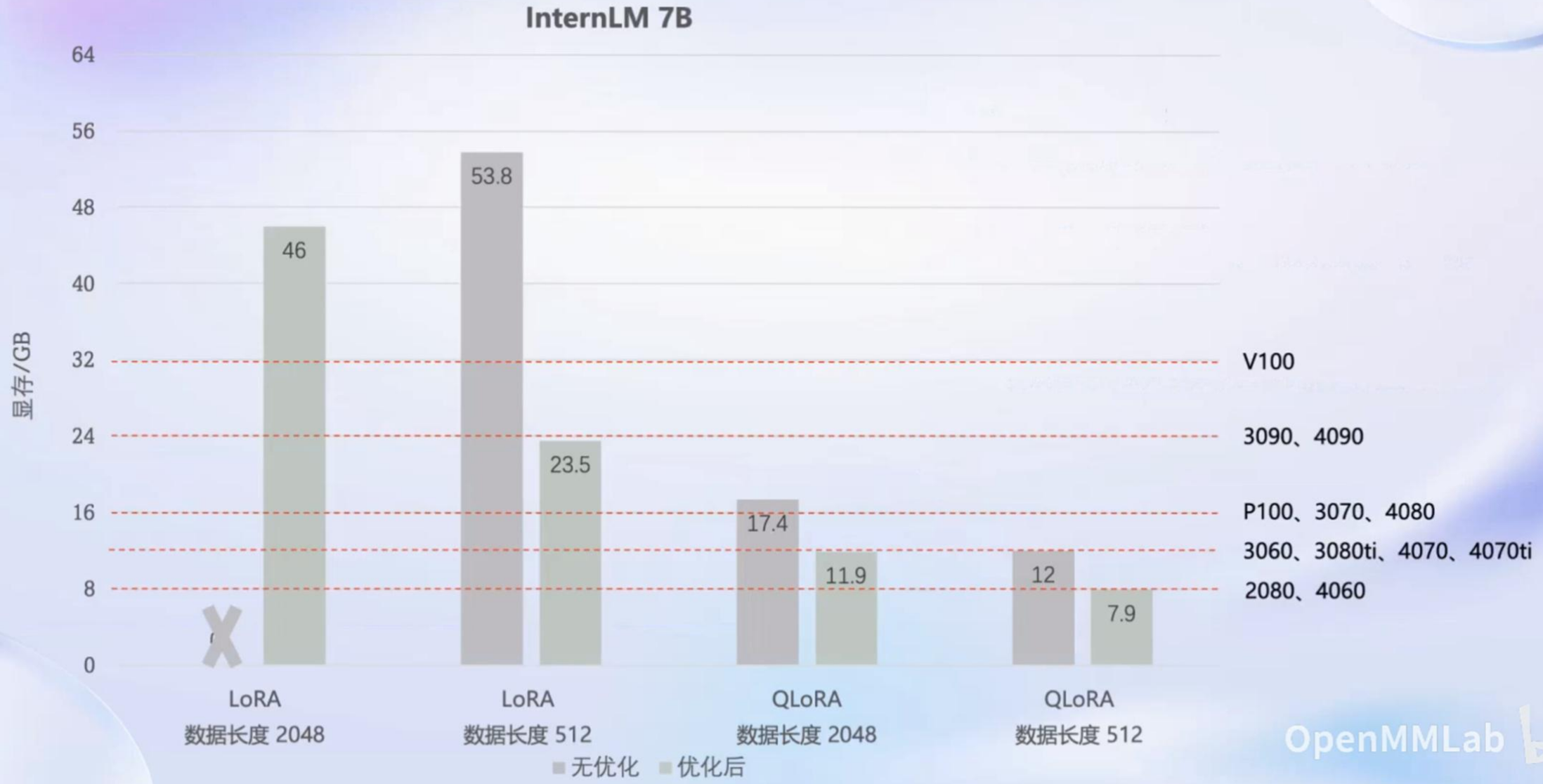
  开发者无需关注复杂的显存优化与计算加速细节

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**

全链条开源开放体系 | 微调

极致的显存优化：消费级显卡（8GB）玩转微调



全链条开源开放体系｜评测

国内外评测体系的整体态势

机构	HELM	FlagEval	MMLU	Alpaca Eval	SuperCLUE 中文大模型综合性测评基准	OpenLLM Leaderboard
机构	Stanford University	北京智源人工智能研究院 BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE	Berkeley UNIVERSITY OF CALIFORNIA	Stanford University	CLUE	Hugging Face
类型	客观评测	客观/主观评测	客观评测	主观评测	客观/主观评测	客观评测
量级	5W+ 英文题目	8W+ 中英双语	1W+ 英文题目	1K+ 英文题目	3K+ 中文题目	2W+ 英文题目

全链条开源开放体系 | 评测

OpenCompass

全球领先的大模型开源评测体系

6大维度，80+评测集，40万+评测题目

学科	语言	知识	理解	推理	安全
初中考试	字词释义	知识问答	阅读理解	因果推理	偏见 有害性
中国高考	成语习语	多语种知识问答	内容分析	常识推理	公平性 隐私性
大学考试	语义相似		内容总结	代码推理	真实性 合法性
语言能力考试	指代消解			数学推理	
职业资格考试	翻译				

全链条开源开放体系 | 评测

OpenCompass 开源评测平台架构

工具层

分布式评测 | 提示词工程 | 评测数据库上报 | 评测榜单发布 | 评测报告生成

方法层

自动化客观评测 | 基于模型辅助的主观评测 | 基于人类反馈的主观评测

能力层

通用能力：学科 | 语言 | 知识 | 理解 | 推理 | 安全

特色能力：长文本 | 代码 | 工具 | 知识增强

模型层

基座模型 | 对话模型

全链条开源开放体系 | 评测

丰富模型支持

开源模型、API
模型一站式评测

分布式高效评测

支持千亿参数模型在海量
数据集上分布式评测

便捷的数据集接口

支持社区用户根据自身需求
快速添加自定义数据集

敏捷的能力迭代

每周更新大模型能力版单，
每月提升评测工具能力

	模型	发布日期	参数量	综合	学科	语言	知识	理解
1	GPT-4 OpenAI	2023/3/15	N/A	72.1	77.2	62	73.5	70
2	ChatGPT OpenAI	2023/3/1	N/A	61.8	62.7	48.6	64.5	64.6
3	WeMix-LLaMA2-70B Shanghai AI Lab	2023/9/13	70B	58.6	62.3	52.6	69	62.9
4	StableBeluga2 Stability AI	2023/7/21	70B	58.1	61.7	50.7	62.2	60.3
5	Qwen-7B Alibaba	2023/8/3	7B	57.6	62.8	52.4	51.4	67.7
6	LLaMA-2-70B Meta	2023/7/19	70B	57.4	57.3	51.6	67.7	60.8
7	LLaMA-2-70B-Chat Meta	2023/7/19	70B	56.1	54.1	48.2	65	62.1
8	InternLM-Chat-7B-8K Shanghai AI Lab & SenseTime	2023/7/6	7B	55.6	56.7	50.4	50.1	66.9

用户遍及国内外知名企业与科研机构

阿里云 aliyun.com
百川智能 BAICHUAN AI
Baidu 百度
智谱·AI
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
小红书
度小满
oppo
NUS National University of Singapore
Tencent 腾讯
XVERSE 元象
美团
Microsoft
UNIVERSITY OF CAMBRIDGE

全链条开源开放体系｜评测

丰富的模型支持

复旦大学 MOSS

Microsoft WizardLM

清华大学 Tsinghua University 智谱·AI ChatGLM ChatGLM2

ANTHROPIC Claude

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory 商汤 sensetime InternLM

OpenAI ChatGPT GPT-4

Carnegie Mellon University Berkeley UNIVERSITY OF CALIFORNIA Vicuna

Stanford University Alpaca

BAICHUAN Baichuan

TII Technology Innovation Institute Falcon

LLaMA

mosaicML MPT

全链条开源开放体系 | 部署

大语言模型特点

内存开销巨大

- 庞大的参数量
- 采用自回归生成token，需要缓存k/v

动态Shape

- 请求数不固定
- token逐个生成，且数量不定

模型结构相对简单

- transformer 结构，大部分是decoder-only

技术挑战

设备

- 低存储设备（消费级显卡、移动端等）如何部署?

推理

- 如何加速 token 的生成速度
- 如何解决动态shape，让推理可以不间断
- 如何有效管理和利用内存

服务

- 提升系统整体吞吐量
- 降低请求的平均响应时间

部署方案

技术点

- 模型并行
- 低比特量化
- Attention优化
- 计算和访存优化
- Continuous Batching

全链条开源开放体系 | 部署

LMDeploy

LMDeploy 提供大模型在GPU上部署的全流程解决方案，包括模型轻量化、推理和服务。

接口

Python | gRPC | RESTful

轻量化

4bit权重

8bit k/v

推理引擎

turbomind

pytorch

服务

openai-server | gradio

triton inference server

高效推理引擎

- 持续批处理技巧
- 深度优化的低比特计算kernel
- 模型并行
- 高效的k/v缓存管理机制

完备易用的工具链

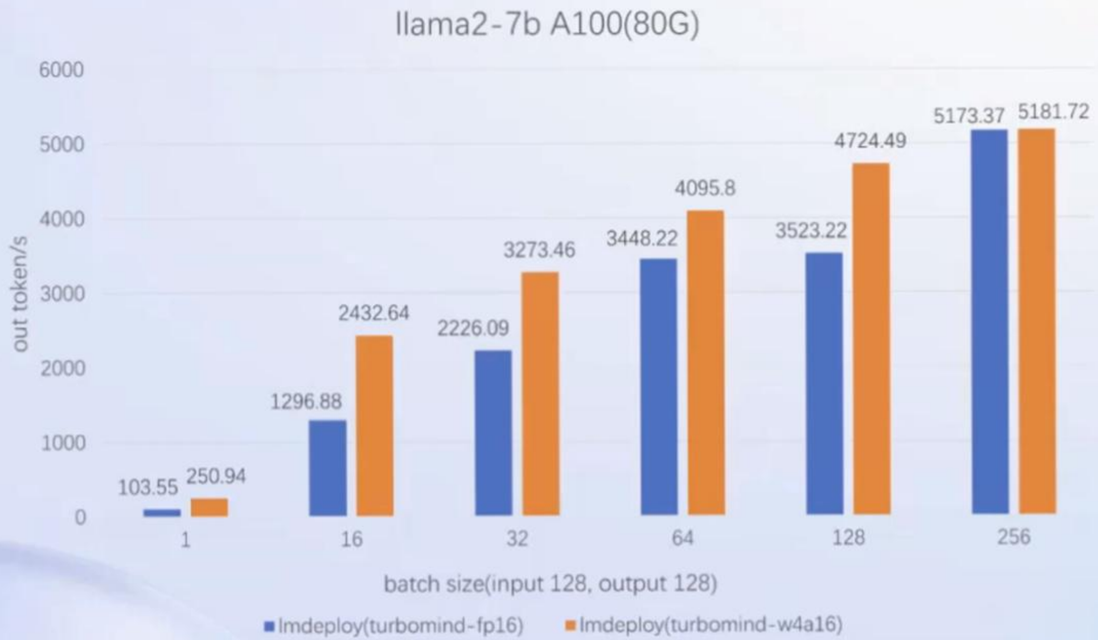
- 量化、推理、服务全流程
- 无缝对接OpenCompass评测推理精度
- 和 OpenAI 接口高度兼容的 API server

全链条开源开放体系 | 部署

领先的推理性能

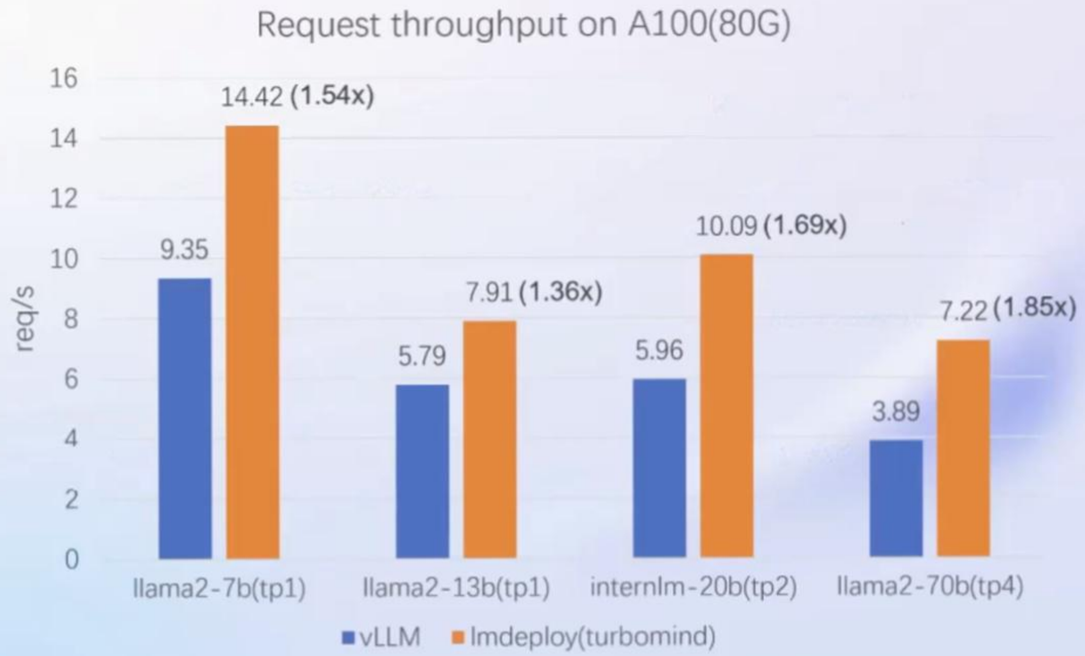
静态推理性能

固定 batch, 输入/输出 token 数量



动态推理性能

真实对话，不定长的输入/输出

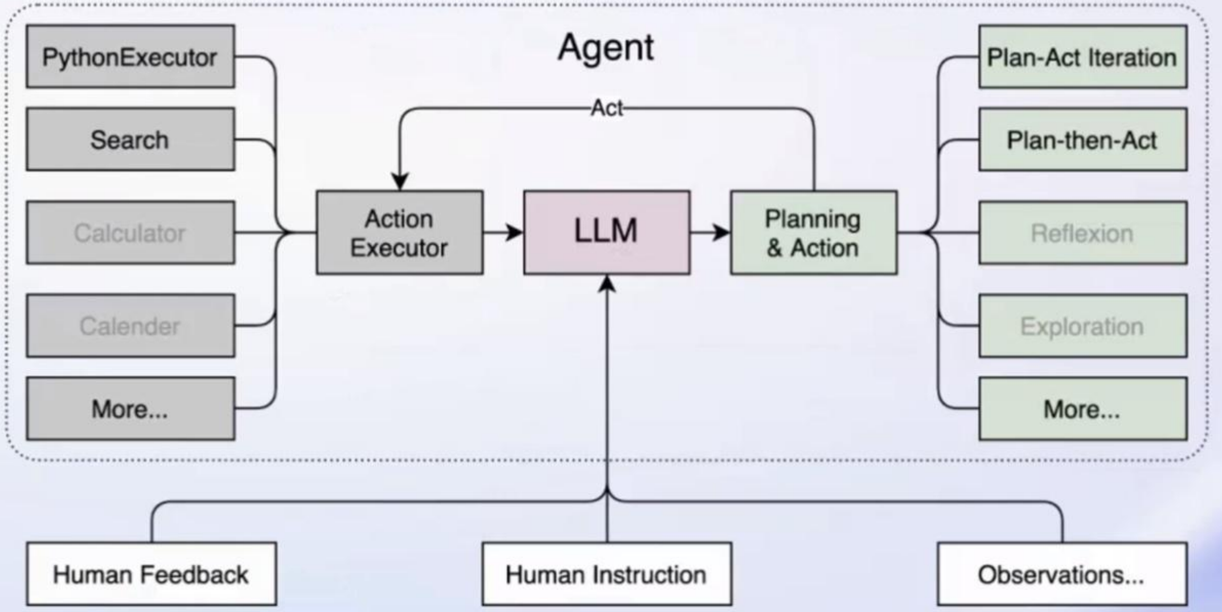


全链条开源开放体系 | 智能体

大语言模型的局限性

- 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

LLM 〉 智能体



全链条开源开放体系 | 智能体

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

GPT-3.5/4

InternLM

Hugging Face Transformers

Llama

简单易拓展，支持丰富的工具

AI 工具	能力拓展	Rapid API
文生图	搜索	出行 API
文生语音	计算器	财经 API
图片描述	代码解释器	体育资讯 API

全链条开源开放体系｜智能体

代码解数学题

设数列
a_n是以1为首项，2为公差的等差数列，数列
b_n是以1为首项，2为公比的等比数列，则$a_1b_1 + a_2b_2 + \ldots + a_{10}b_{10} =$ ____ .

LLM

Python

思 考 步 骤: 由题目条件结合数列知识可以得出，a_n和b_n的通项公式，从而计算出最终结果

插　　件: python

执 行 内 容:

```python
from sympy import *

def solution():
    a,b,n = symbols('a,b,n',real=True)

    # 求解数列 a 的通项公式
    a_n = 1 + 2 * (n-1)
    a_n_10 = a_n.subs(n, 10)

    # 求解数列 b 的通项公式
    b_n = 2**(n-1)
    b_n_10 = b_n.subs(n, 10)

    # 计算答案
    result = a_n_10 * b_n_10
    for i in range(1, 10):
        result += (a_n.subs(n, i) * b_n.subs(n, i))

    return {'result': result}
```

执 行 结 果: {'result': 17411}

17411

重新生成

零样本泛化：多模态 AI 工具使用

Lagent Web Demo

please describe the image and save to audio

Uploaded image

File saved at: /mnt/hwfile/llmit/code/lagent/tmp_dir/e039001d-df05-4c6c-903e-6456570f75f7.png

Image Description

插　　件: Image Description

思 考 步 骤: I need to use the Image Description and Text Reader APIs to describe the image and save it to audio.

执 行 结 果:

a puppy and a cat sitting on a blanket

Text Reader

插　　件: Text Reader

思 考 步 骤: Now that I have the description of the image, I can use the Text Reader API to save it to audio.

执 行 结 果:

/mnt/hwfile/llmit/code/lagent/generated/audio/20230918-3ee3.wav

0:03 / 0:03

FinishAction

The image you uploaded is of a puppy and a cat sitting on a blanket. You can listen to the audio description of the image by visiting the following link: /mnt/hwfile/llmit/code/lagent/generated/audio/20230918-3ee3.wav.

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

书生·浦语全链条开源开放体系

数据

书生·万卷

2TB数据，
涵盖多种模态与任务

预训练

InternLM-Train

并行训练，极致优化
速度达到 3600 tokens/sec/gpu

微调

XTuner

支持 全参数微调，
支持LoRA等低成本微调

部署

LMDeploy

全链路部署，性能领先
每秒生成 2000+ tokens

评测

OpenCompass

全方位评测，性能可复现
80 套评测集，40 万道题目

应用

Lagent
AgentLego

支持多种智能体，支持代码解释器等多种工具

谢谢

https://www.shlab.org.cn
Shanghai Artificial Intelligence Laboratory

OpenMMLab