

(1)书生·浦语大模型全链路开源体系

课程链接：【(1)书生·浦语大模型全链路开源体系】

<https://www.bilibili.com/video/BV1Rc411b7ns/>

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（围棋 AlphaGo，蛋白质折叠 AlphaFold，语音识别，人脸识别等）

通用模型：一个模型应对多种任务，多种模态。（GPT4，PaLI）

书生·浦语大模型开源历程



从大语言模型到多模态到智能体到工具链全线升级

书生·浦语大模型性能简介

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

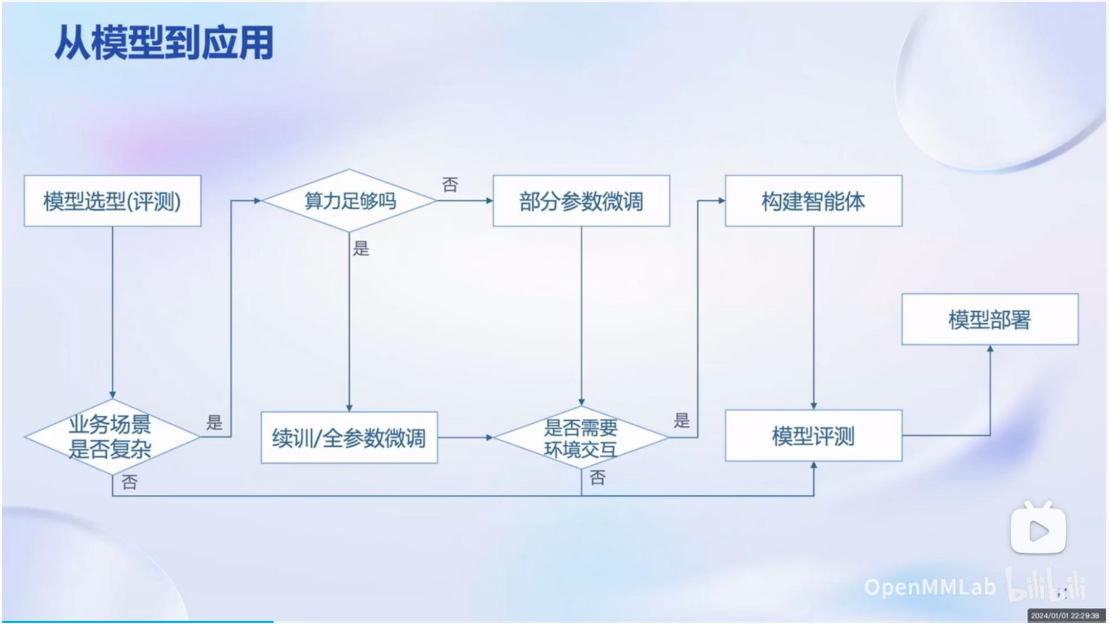
社区低成本可用最佳模型规模 商业场景可开发定制高精度较小模型规模 通用大语言模型能力全面覆盖

05:41 / 42:00

轻量级和中量级重量级的区分标准是什么，为什么分别是这么多参数？几个不同量级的模型本质是 transformer 层数不同吗？

性能上，全面领先相近量级的开源模型，Internlm-20b 以不足 1/3 的参数量接近 llama2-7B 的水平。

从模型到应用



书生·浦语全链条开源体系



从数据获取到模型应用每个环节都有开源

大模型的未来发展思考

大模型的未来是要更方便的深入到日常生活中，所以轻量级的部署和更快的推理速度是必须追求的。

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



OpenMMLab

2024/10/10 22:49:51

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

llama2-7b A100(80G)



batch size(input 128, output 128)	lmdeploy(turbomind-fp16)	lmdeploy(turbomind-wfa16)
1	103.55	250.94
16	1296.88	2432.64
32	2226.09	3273.46
64	3448.22	4095.8
128	3523.22	4734.49
256	5173.37	5181.72

动态推理性能

真实对话，不定长的输入/输出

Request throughput on A100(80G)



Model	vLLM	lmdeploy(turbomind)	Speedup
llama2-7b(tp1)	9.35	14.42	1.54x
llama2-13b(tp1)	5.79	7.91	1.36x
internlm-20b(tp2)	5.96	10.09	1.69x
llama2-70b(tp4)	3.89	7.22	1.85x



OpenMMLab

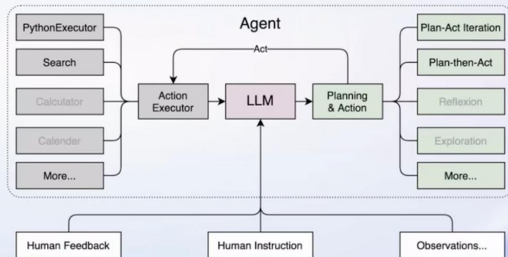
2024/10/10 22:54:17

通过大模型代替人与世界交互，才能充分解放人类，让人从事创造性的生产活动。于是有了智能体 agent

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

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

